DATA DRIVEN APPROACHES FOR DISSECTING TUMOR HETEROGENEITY

by

ARDA DURMAZ

Submitted in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

Department of Nutrition

Systems Biology and Bioinformatics

CASE WESTERN RESERVE UNIVERSITY

January 2023

CASE WESTERN RESERVE UNIVERSITY

SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis of

ARDA DURMAZ

Candidate for the **Doctor of Philosophy degree***.

Committee Chair

David T. Lodowski PhD.

Committee Member

Peter C. Scacheri PhD.

Committee Member

Satish E. Viswanath PhD.

Committee Member

Jacob G. Scott, MD, DPhil

Date: December 7, 2022

*We also certify that written approval has been obtained for any proprietary material contained therein.

Dedicated to my parents
Timur DURMAZ and Şebnem DURMAZ
for their unwavering support and guidance

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

"The highest forms of understanding we can achieve are laughter and human compassion."

– Richard P. Feynman

First, I would like to thank *Jacob G. Scott MD. DPhil.*, my thesis advisor, for his guidance and support. Thank you for showing me the humility of not knowing and the excitement of finding things out. Thank you for cultivating such a diverse, welcoming and intellectual incubator. Secondly, I want to thank *Jaroslaw Maciejewski MD. PhD.*, *Valeria Visconte PhD.* and *Carmelo Gurnari MD. PhD.* for their guidance, mentorship and spirited discussions that eventually lead to a body of research presented in this work.

I also want to thank the members of Theory Division for their day-to-day unwavering, contagious curiosity and to the past and present members of my committee, *Peter C. Scacheri PhD., David T. Lodowski PhD., Satish E. Viswanath PhD. and Tae Hyun Hwang PhD.* for their encouragement and feedback.

Finally, I am forever thankful to my wife, *Ayça*, for always being there for me in my journey and for always showing me the value of just being. Without you, I could not have crossed the seas.

Data Driven Approaches for Dissecting Tumor Heterogeneity

Abstract

by

ARDA DURMAZ

Molecular heterogeneity in cancer has been recognized as one of the main drivers of disease relapse and drug resistance. In addition, effects of tumor-microenvironment have shown to contribute to the diversity as well. Consequently, cancer research has aimed at generating and utilizing inherently high-dimensional molecular datasets for the past decade to characterize tumors specifically with the development of 'sequencing-by-synthesis' Next-Generation Sequencing (NGS) platforms. Large collections of high-dimensional multi-omics datasets exemplified by TCGA and PCAWG, 1) elaborate on the heterogeneity of cancer progression and 2) allow for increasingly complex models to be utilized. Respecting the black-box nature of machine-learning driven models, here we develop multiple strategies to leverage molecular information to delineate disease progression/mechanisms in Leukemia. Furthermore, we show the requirement of careful selection of strategies in noisy scRNA-Seq datasets in solid tumors and we propose an integrative model to investigate collateral drug-responses in a pan-cancer fashion.

We present multiple strategies in two parts. First, we present a *Bayesian Latent Class Analysis* to incorporate molecular information in a large cohort of (n = 2681) AML patients with heterogeneous characteristic and generate novel unsupervised clusters with clinical relevance. Furthermore, we utilize Autoencoder structure to develop distance-based, low dimensional clustering model to group MDS patients (n = 3588) into 14 novel

groups. This approach allowed us to extract relevant features otherwise difficult to capture with Bayesian strategies in noisy datasets. In the second part, we conduct a comprehensive benchmarking study to evaluate the vast repository of methods developed for scRNA-Seq analysis. We show, in contrast with the current practice, scRNA-Seq analysis is amenable to variation and results, specifically unsupervised clustering, is of qualitative nature rather than quantitative. Finally, borrowing from the power of complex neural-network based models, we develop an integrative model to capture co-varying features of gene-expression and mutation profiles of cell-lines and patient samples relevant to collateral drug response profiles.

Chapter 1

Introduction & Motivation

Chapter 1

INTRODUCTION & MOTIVATION

The inherent complexity of cancer both as a population of cells and as individual cells results in cancer's fascinating ability to adapt to environmental stress whether in the form of resistance evolution, immune evasion or radiation resistance. In order to better understand and consequently devise better treatment regimens and control the adaptation process, models ranging from purely theoretical to purely data-driven have been developed. However, the heterogeneity of tumors, genetic and non-genetic, have hindered a one size fits all solution, specifically in high-grade tumors both for solid tumors and leukemias. Single-cell sequencing technologies elaborated further on the 'uniqueness' of cancer and identified novel mechanisms/cell populations driving this observed heterogeneity. This issue is further exacerbated by the high-dimensionality of biological processes. Fortunately, the exponentially increasing availability of data in repositories and progress in the machine-learning (ML) models have made cancer research amenable to increasingly complex integrative models. Herein we apply multiple ML strategies in different settings and cancers to integrate these high-dimensional heterogeneous processes to clinically amenable features. In part 1, for Leukemia, we utilize sparse mutation profiles obtained from whole-exome sequencing as binary inputs to cluster the patients in an unsupervised fashion via both a model based and distance based approach (**Fig.**1.1). In contrast, in part 2, for solid tumors we use single-cell RNA sequencing to model the variability in combinatorial analysis followed by multi-omics integration for drug sensitivity prediction in a pan-cancer fashion using whole-transcriptome and whole-exome datasets (**Fig.**1.2).

1.1 Current State of Genomic Classification in Myeloid Neoplasms

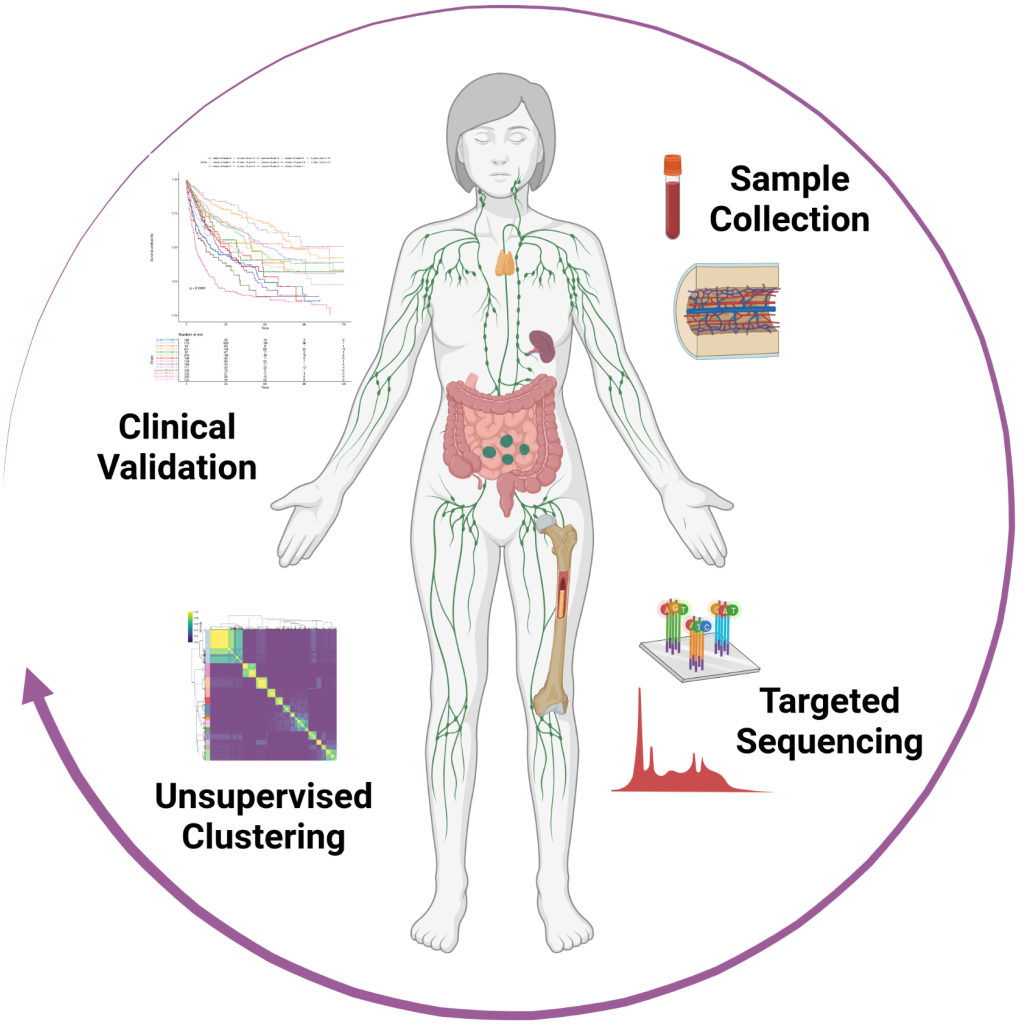


Figure 1.1: **Datasets and analysis steps used in Leukemia**. Peripheral blood and or bone-marrow samples were in a targeted fashion. Obtained mutation profiles are binarized based on frequency and quality threshold. Using these mutation profiles, unsupervised clustering via consensus approach followed by clinical validation by time-to-event modeling was performed.

Myeloid Neoplasms (MN) are a complex group of clonal diseases of the haematopoietic system primarily characterized by morphological features and cytogenetic abnormalities. Whole exome and genome sequencing studies have been performed for about a decade and have shown a much higher level of mutational complexity compared to what was previously known. Hotspot mutations in components of the RNA-splicing machinery

(among the most frequent: PRPF8, SF3B1, SRSF2, U2AF1, ZRSR2) have been discovered; however discovery of these mutations have not yet translated in clinical applicability. In fact pharmacologic agents directed against such hotspot have failed in experimental clinical trials [1, 2].

A large number of studies have instead focused on the description of mutations in master transcription factors regulating the differentiation process from hematopoietic stem cells (HSCs) to common myeloid/lymphoid progenitors (CMP/CLP) such as *RUNX1* and *GATA2*. However, the characterization of disease subtypes is complicated by possible clinically relevant interactions in addition to the continuous nature of these clinical features. For instance, blast percentage cut points have been used to distinguish myelodysplastic syndrome (MDS) from acute myeloid leukemia (AML) with a threshold of 20% myeloblasts in bone marrow or peripheral blood [3]. Similarly in chronic myeloid leukemia (CML) which is primarily characterized by the BCR-ABL1 fusion, is further stratified as progressive/blast phase (BP) depending on $\geq$ 20% myeloid blasts in blood or bone marrow, or the presence of extramedullary proliferation of blasts where 20% cut-off is sub-optimally defined to represent the BP. However a fixed blast percentage to distinguish both diseases is still a debate in clinical practice.

In order to formally define distinct disease entities and classify patients according to diverse cytogenetic and routine parameters, international efforts to build consistent scoring systems have been deployed. This is the case of The European LeukemiaNet [4, 5] risk stratification in Acute Myeloid Leukemia (AML), IPSS-R [6] which is further improved by incorporating molecular information IPSS-M [7] in MDS. These efforts have been made towards to re-definition of the World Health Organization (WHO) classification. In these new and latest models, machine learning (ML) based unsupervised approaches have been incorporated [3, 8–10]. However, the heterogeneous nature of MN can lead to variability across the studies regarding the subtypes identified dependent on the characteristic of the cohort, number of patients and available genomic information. Hence robust, unsupervised, integrative

approaches are required to delineate disease progression.

Here, we first briefly describe MN classifications currently adopted to further elaborate the disease heterogeneity and then describe the ML approaches to 1) group AML patients based on molecular features in a model driven framework and 2) using unsupervised dimension-reduction to group MDS patients in an unbiased fashion.

1.1.1 Myeloproliferative Neoplasms

Myeloproliferative neoplasms (MPN) result in an increased production of functional terminal hematopoietic cells and can be broadly categorized under 2 distinct classes; The *BCR-ABL1*$^{+}$ fusion which characterizes CML and the *BCR-ABL1*$^{-}$ cases which can be further stratified into seven distinct disease entities; polycythaemia vera (PV), essential thrombocythaemia (ET), primary myelofibriosis (PMF), chronic neutrophilic Leukemia (CNL), chronic eosinophilic leukemia (CEL), juvenile myelomonocytic leukemia (JMML) and not otherwise specified (NOS).

As said above, CML is driven by the *BCR-ABL1*$^{+}$ fusion gene encoding an oncogene with kinase activity which through multiple mechanisms leads to the transcription of *BCL2* and *BCL2L1*, increasing anti-apoptotic mechanisms [11]. Consequently, Imatinib, a tyrosine kinase inhibitor (TKI) has been shown to be highly effective with a median overall survival of 11 years [12]. In contrast, *BCR-ABL1*$^{-}$ MPN stratification is largely characterized by morphological criteria. For instance, major criteria for PV include hemoglobin $\geq 16.5 g/dL$ in males and $\geq 16.0 g/dL$ in females.

Similarly, ET requires bone marrow biopsy to differentiate from prePMF [13]. Nevertheless, molecular studies have identified polymorphisms that improve stratification for patients with variants; *JAK2, CALR, MPL* shared across PV, ET and PMF. Similarly, majority of CNL patients carry *CSF3R* mutations for which the diagnosis is made based on white-blood cell count (WBC) $\geq 25 \times 10^{9}/L$.

1.1.2 Myelodysplastic Neoplasms

In contrast to MPN, Myelodysplastic Neoplasms (MDS) are characterized by a decrease in production of hematopoietic cell lineages and dysplasia. Current classification of MDS stratifies into: 1) MDS with defining genetic abnormalities and 2) MDS, morphologically defined. Genetic abnormalities include several mutations (among the most frequent: *SF3B1* and *TP53*), cytogenetic abnormalities include 5q, 7q and 20q deletions, monosomy 7 and complex karyotype.

In combination with morphological features, genomic alterations characterizing MDS cases such as; MDS-*SF3B1* which presents low blast counts and *SF3B1* mutations (See [3] for a full description of the updated classification). Further characterization of shared features between MDS and MPN have been defined as well. A major classification of MDS/MPN group includes chronic myelomonocytic leukemia (CMML) which is characterized by consistently high levels of monocytes in blood with or without cognate genetic abnormalities such as splicing inefficiencies and epigenetic regulation. To further complicate the distinctions, 2 subtypes for CMML has been defined; myelodysplastic CMML and myeloproliferative CMML based on decreased or increased levels of WBC respectively (with the threshold for WBC $13 \times 10^9/L$).

1.1.3 AML

AML, similar to MPN and MDS manifests a heterogeneous etiology, however the distinction between MDS and AML is rather arbitrary. Previously defined as $\geq$ 20% blasts, current understanding has shifted towards a more covariate inclusive criteria.

Due to advances in treatment options, for patients younger than 60 years, chemotherapeutics are among the most effective options available. For patients unable to receive intensive chemotherapy generally older patients treatment options are lacking. Furthermore, advancements in molecular feature screening have prompted for revisions of AML classifications previously focused on cytogenetics such as *PML-RARA,CBF-MYH11* rearrangements by

incorporating single-nucleotide variations in *FLT3, NPM1, CEBPA, KIT* genes which can be observed in mutually-exclusive fashion with cytogenetic abnormalities. Broadly, AML can be categorized, based on disease etiology into 4 distinct groups; *de-novo*/primary AML associated with genetic abnormalities, secondary AML associated with a previous MDS related condition, and therapy related AML due to a prior cytotoxic treatment and Not Otherwise Specified (NOS/Defined by Differentiation) (See [5] for further details).

De-novo AML is characterized by, as given above, defined cytogenetic abnormalities such as fusion-genes and mutations with or without additional morphological features such as blast percentage. AML NOS or currently updated definition AML defined by differentiation, is characterized by both morphological features such as low blasts and absence of differentiation markers such as *CD13, CD33* and *CD117*. Secondary AML on the other hand is characterized by either the presence of cytogenetic abnormalities associated with MDS and/or MPN or the preexistence of myeloid disorder. Similarly, therapy related AML includes cases previously treated with cytotoxic agents and associated with high frequency of *TP53* mutations.

Efforts to guide therapy have also characterized patients in terms of risk profiles. For instance, widely accepted ELN risk stratification includes three groups; Favorable, Intermediate and Adverse based on the cytogenetic and mutation profiles [4]. However, further efforts to update risk stratification in an unbiased and comprehensive manner suggested changes to defined groups as well

In order to alleviate some of the issues present in stratification of MN patients into clinically relevant subgroups, we utilized unsupervised ML approaches enabling us to combine the heterogeneous molecular spectra modeling frequency of co-occurrence of mutations. We first utilize Latent Class Analysis to extract patient clusters and show that the identified clusters further elaborate the mixed disease defining entities where primary AML and secondary AML cases can cluster together showing similar molecular profiles with distinct

pathophysiological progression. Secondly, we adopt a dimension-reduction approach to embed MDS patients onto a low-dimensional manifold based on the similarity of mutation profiles effectively capturing the covariance structure of the features in the patient cohort. We identify 14 clusters that show distinct clinical characteristics. The identified 14 clusters are further aggregated into 5 risk groups that show significant time-to-event profiles. We provide these ML models as an open-source tool for the community and hope that the approaches presented here, and further improvements, will pave the way for unbiased and robust exploration of molecular features allowing for better treatment opportunities.

1.2 Heterogeneity in Solid Tumors Elucidated through Single-Cell RNA Sequencing

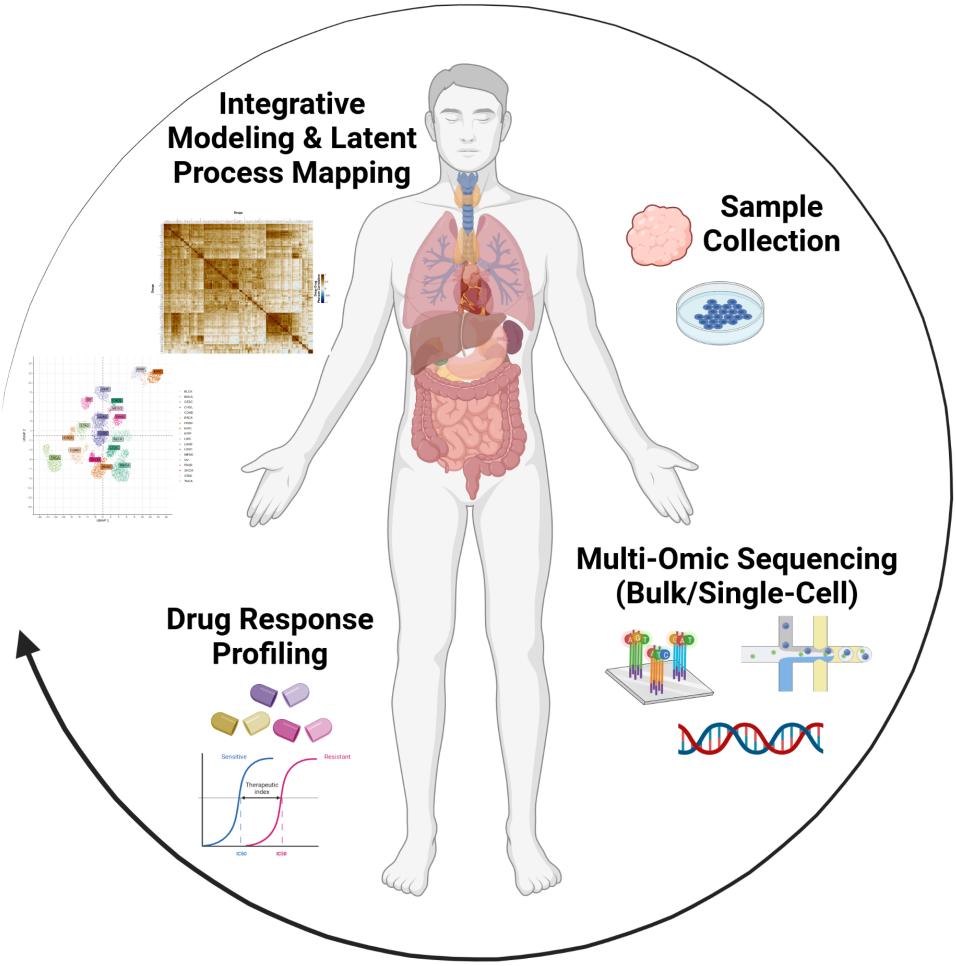


Figure 1.2: **Datasets and analysis steps used in solid tumors**. We obtained multi-omics datasets from public repositories (*e.g.* GDSC, TCGA) in addition to an in-house generated single-cell RNA-Seq (scRNA-Seq) dataset. First, using the scRNA-Seq datasets, we evaluated multiple combinatorial workflows of dimension reduction, clustering and pseudotime/trajectory mapping. Secondly, integrating the multi-omics data of gene-expression, whole exome sequencing and drug sensitivity profiles, we identified collateral signatures in a pan-cancer fashion.

Solid tumors present an increased dysregulation in homeostatic processes where cells utilize both genetic and epigenetic mechanisms to proliferate, migrate and evolve in a stochastic and Darwinian fashion. Stochasticity results in increased intra-tumoral heterogeneity (ITH) in mechanism and spatial organisation, and this apparent ITH can be further stratified into

genetic and non-genetic components including chromosomal aberrations and transcriptional regulations respectively [14]. Furthermore, ITH has been previously associated with increased propensity for both intrinsic and evolved resistance to treatment which is non-trivial to study using bulk sequencing technologies [15, 16].

Relatively recently, single-cell sequencing technologies have been adapted to better study ITH and probe transcriptional dynamics at single-cell resolution. Specifically, scRNA-Seq analysis allowed for capturing distinct sub-populations via clustering or for ordering cells on a latent transcriptional process termed pseudotime ordering or trajectory analysis.

Although scRNA-Seq is one of the first widely used method, alternative technologies for epigenomic and proteomic quantification methods have also been developed [17–19]. Going further, single-cell sequencing methods are being developed to profile multi-modal data as well [20–22]. However, capture of individual cells and individual molecules is not without challenges. Specifically, for scRNA-Seq, low transcriptional coverage due to dropouts, and reduced total depth due to increased number of cells results in noisy and sparse datasets which require subsequent imputation and robust normalization methods to account for technical noise.

Multiple approaches have been developed to address technical challenges in scRNA-Seq analysis including the explicit modeling of technical and biological noise [23], clustering based on dropouts [24] and several imputation [25, 26] and robust normalization methods [27–29]. Dimension reduction techniques have been extensively adapted to scRNA-Seq studies as well with the aim of capturing dominant transcriptional patterns hence effectively reducing technical noise.

Both linear dimension reduction techniques such as Principal Component Analysis, Metric Multi-Dimensional Scaling and non-linear dimension reduction methods such as t-Stochastic Neighbor Embedding [30] and Uniform Manifold Approximation and Projection [31] are used. Dimension reduction is further coupled with unsupervised clustering

methods such as k-means and density based alternatives (DBSCAN) [32] in order to discover distinct cell populations based on transcriptional activity. Furthermore, the selection of which method to adapt and couple with downstream analysis is seldom made with combinatorial nature of the data in consideration. This issue has been, in the context of trajectory analysis, evaluated partially in relatively homogeneous models.

In order to further elaborate on the use of different workflows adapted in scRNA-Seq analysis in a combinatorial fashion, specifically in the context of but not limited to the evolution of drug resistance, we evaluated over 6k analysis combination in the context of unsupervised dimension reduction followed by subsequent clustering and trajectory analysis. We showed that it is non-trivial to develop repeatable workflows especially when the transcriptome coverage and/or number of cells captured is relatively low. Furthermore, we showed that regularization in the context of dimension reduction improves repeatability in identified clusters. In the context of trajectory identification however, reduced overlap of identified pseudo-ordering of cells on a latent process whether it is due to drug induced selection or developmental process can be alleviated by methods that can take into account temporal information. Overall, we hope that the results presented can guide researchers utilizing scRNA-Seq datasets and ameliorate current reproducibility issues due to arbitrary selection of parameters and tools.

1.3 Integrative Modeling of Multi-omics Data to Characterize the Drug Sensitivity Landscape

Cancer treatment is hindered by evolution and the emergence of cell populations that can proliferate under environmental stress. The diversity of mechanisms individual cells can 'rewire' in order to survive is exacerbated by the inherent stochasticity of cell states which requires better treatment strategies to overcome drug resistance. Adaptive therapies, drug holidays, metronomic treatment are such strategies aiming to account for evolutionary dynamics prolonging overall survival and/or resistance evolution [34, 35].

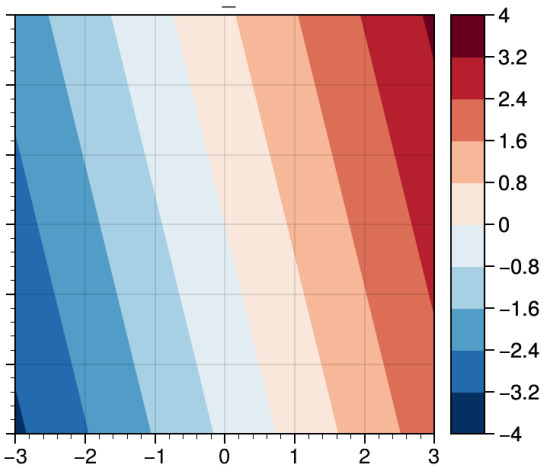


Figure 1.3: **Example linear fitness landscape showing the association of reduced dimensions with the phenotype of interest**. We generated 2d representation of binary mutation profiles of *E.coli* supervised by the growth-rate (fitness) measurements under Cefazolin treatment using partial-least squares simplifying the high-dimensional data into a more tractable representation [33].

Furthermore, efforts to understand the underlying dynamics, led to the development of collateral response models where collateral-sensitivity is defined as the increased sensitivity to a drug due to the increased 'cost' of developing resistance to another drug. Multitudes of studies have generated networks of collaterally sensitive or resistant drug pairs both in bacteria and in cancer showing the sparse and heterogeneous nature of such paired mechanisms [33, 36–39].

In parallel, genotype-fitness maps which were first defined by Wright *et al.*[40] later generalized by Kauffman *et al.*[41] are utilized to study tumorigenesis. These maps/landscapes are

static functions mapping from genotype space to fitness space where temporal information can be integrated to capture the dynamic nature (*e.g. drug concentration*) termed fitness seascapes [42].

The utility of these functions manifest through 'simplification' of high-dimensional genotype space and effectively representing the evolutionary process in low dimensional space making it tractable (See Fig. 1.3). Consequently, genotype-fitness maps/landscapes can be characterized via manifold-learning or low-dimensional embedding methods in a supervised fashion. A straightforward example would be the eigendecomposition of the covariance matrix of variables (*e.g* protein abundance) and fitness (*e.g* drug sensitivity) for instance. Manifold-learning has been the de-facto approach for high-dimensional datasets and several methods have been developed and in standard practice (*e.g Principal Component Analysis and t-SNE*).

We hypothesize that the genotype-fitness maps/landscape framework can be effective in characterizing the collateral mechanisms of drug sensitivity through low-dimensional embedding in a supervised fashion. In this section, we investigate the utility of integrative machine-learning approach based on autoencoders in determining the collateral sensitivity potential of given pair of drugs based on coupled sensitivity signatures. We utilize multi-omics datasets consisting of gene expression, mutation and drug sensitivies in a pan-cancer fashion which can be formally defined as joint dimensionality reduction (jDR) [43]. We use datasets from Genomics of Drug Sensitivity in Cancer (GDSC), Dependency Map (DepMap) and The Cancer Genome Atlas (TCGA) [44, 45]. As a proof of concept, we show the capability of such models in integrating multiple -omics datasets to uncover potential convergent mechanisms of resistance/sensitivity in a pancancer fashion. We further recapitulate the sparse nature of collaterally sensitive drug pairs. We aim to expand the proposed model to scRNA-Seq data as well.

PART I

Data Driven Approaches in Leukemia

- Published as: Awada, H., **Durmaz, A**., ... Maciejewski, J. P. (2021). Machine learning integrates genomic signatures for subclassification beyond primary and secondary acute myeloid leukemia. Blood, 138(19), 1885-1895
- In Review as: Kewan, T., **Durmaz, A**., ... Visconte, V., Maciejewski, J., P., Molecular patterns identify distinct subclasses of myeloid neoplasia. [Manuscript Submitted in Nature Communications]

Chapter 2

MACHINE LEARNING INTEGRATES GENOMIC SIGNATURES FOR SUBCLASSIFICATION BEYOND PRIMARY AND SECONDARY ACUTE MYELOID LEUKEMIA

While genomic alterations drive the pathogenesis of acute myeloid leukemia (AML), traditional classifications are largely based on morphology and prototypic genetic founder lesions define only a small proportion of AML patients. The historical subdivision of primary/de novo AML (pAML) and secondary AML (sAML) has shown to variably correlate with genetic patterns. Perhaps, the combinatorial complexity and heterogeneity of AML genomic architecture have precluded, so far, the genomic-based subclassification to identify distinct molecularly-defined subtypes more reflective of shared pathogenesis. We integrated cytogenetic and gene sequencing data from a multicenter cohort of 6,788 AML patients that were analyzed using standard and machine learning methods to generate a novel molecular subclassification of AML with biological correlates corresponding to underlying pathogenesis. Standard supervised analyses resulted in modest cross-validation accuracy when attempting to use molecular patterns to predict traditional pathomorphological AML classifications. We performed unsupervised analysis by applying Bayesian Latent Class method that identified 4 unique genomic clusters of distinct prognoses. Invariant genomic features driving each cluster were extracted and resulted in 97% cross-validation accuracy when used for genomic subclassification. Subclasses of AML defined by molecular signatures overlapped current pathomorphological and clinically-defined AML subtypes. We internally and externally validated our results and share an open-access molecular classification scheme for AML patients. Hence, although the heterogeneity inherent in the genomic changes across nearly 7,000 AML patients is too vast for traditional prediction methods, however, machine learning methods allowed for the definition of new genomic AML sub-

classes indicating that traditional pathomorphological definitions may be less reflective of overlapping pathogenesis.

2.1 Introduction

Genetic mutations (somatic or germline), cytogenetic abnormalities and their combinations contribute to the heterogeneity of acute myeloid leukemia (AML) phenotypes [46–48]. Seminal studies have described the molecular landscape of AML and its exclusive framework and have stratified patients into prognostic subgroups [10, 49, 50]. Moreover, serial sequencing studies have delineated a stepwise acquisition of subclonal mutations accompanying AML evolution [51]. To date, prototypic founder lesions [e.g., t(8;21), inv(16), t(15;17)] define only a fraction of AML subgroups with specific prognoses corresponding to molecular pathogenesis [4, 13]. Indeed, in a larger proportion of AML patients, somatic mutations or cytogenetic abnormalities potentially serve as driver lesions in combination with numerous acquired secondary hits [47]. However, their combinatorial complexity hampers the resolution of distinct genomic classifications and overlaps across classical pathomorphological AML subtypes, including de novo/primary (pAML) and secondary AML (sAML) evolving from an antecedent myeloid neoplasm (MN) [52, 53]. These AML subtypes are themselves nonspecific due to variable understanding of their pathogenetic links, especially in cases without overt dysplasia [54, 55]. Without dysplasia, reliance is mainly on anamnestic clinical information that might be unavailable or cannot be correctly assigned due to a short prodromal history of antecedent MN. Additionally, supervised analytical strategies to reproduce current pathomorphological entities as "gold standard" using molecular features have been modest. Here, we explored the potential use of distinct genomic markers, uncovered by advanced machine learning methods, to sub-classify AML objectively and provide personalized prognostication, irrespective of the clinicopathological information, and thus propose to become a standard in AML assessment. We analyzed integrated genomic data from pAML and sAML patients seen in our institution and multiple

other centers over two decades using both standard supervised approaches and unsupervised machine learning methods that better captured the complex interactions of high-dimensional genomic features underlying AML subgroups. Machine learning was instrumental for the identification of novel AML subgroups of invariant driver genomic features.

2.2 Methods

2.2.1 Patients and cell samples

.For the purpose of this study, we combined AML patient data from the Cleveland Clinic (CC, n=855) and the Munich Leukemia Laboratory (MLL, n=4002) with publicly available datasets (The Cancer Genome Atlas, The BEAT AML Master Trial and The German-Austrian Study Group; n=1931, cases with unavailable cytogenetics were excluded)[10, 49, 56] to form a large, well-annotated cohort of 6788 patients (**Table S A1**). Targeted next-generation sequencing (NGS) results, at time of AML diagnosis, were adjusted to focus on the most recurrent somatic myeloid mutations (Table.S2). Patients' follow-up was up to September 2019 with a median duration of 12.4 months. Peripheral blood and/or bone marrow samples were collected after receiving written informed consent according to protocols approved by the Institutional Review Board at CC and other institutions in accordance with the Declaration of Helsinki. Clinical parameters were obtained from medical records after securing appropriate material transfer agreements and from resources accessible online.

2.2.2 Genetic studies.

For the data collected at CC, whole-exome sequencing (WES) was performed on paired tumor and germline DNAs (purified CD3+ lymphocytes). Whole-exome capture was accomplished according to SureSelect Human All Exon 50Mb or V4 kit (Agilent Technologies) and captured targets were sequenced using a HiSeq 2000 (Illumina). Reads were aligned to the human genome (hg19) by a Burrows-Wheeler aligner (http://bio-bwa.sourceforge.net/). Data were validated using a TruSeq Custom Amplicon kit (Illumina) with a panel of 44

genes (**Table.S2**). Variants were annotated using Annovar and filtered and a bio-analytic pipeline developed in-house[57, 58] identified somatic mutations as specified in Supplemental Material. Variants in the patients from the MLL cohort were called as previously reported [59–62]. The gene sequencing methods of publicly-shared AML patients were previously described [10, 49, 56].

2.2.3 Statistical analyses.

Multivariate Cox Proportional-Hazards (Cox-PH) modeling was used to identify genomic abnormalities associated with survival in various AML cohorts. Uni- and multivariate logistic regression (ULR and MLR, respectively) analyses were performed to find distinct genomic features of pAML and sAML. We performed unsupervised analysis to cluster AML patients into genomic subgroups by latent variable modeling. More specifically, we used Bayesian Latent Class Analysis (BLCA) coupled with resampling to generate a consensus-matrix[63] that was then hierarchically clustered using Ward's criteria to obtain final patient cluster assignments. To validate the prognostic significance of identified clusters, we used survival analysis. To determine if AML subtype distributions differed across identified clusters, we normalized pAML and sAML samples to population proportions using bootstrap method. To identify distinct genomic features and generate a subclassification model, we used Random Forest (RF) classification and extracted the variables with the highest global importance measured by mean decrease in accuracy. Additionally, we performed internal and external validation of our model. Finally, the RF subclassification model and cluster-specific survival estimates are available via a web-based open-access resource.

2.3 Results

2.3.1 Molecular architecture determines disease risk and distinguishes AML subtypes

Using the World Health Organization (WHO) 2016 diagnostic criteria[13], we classified 6788 AML patients as core-binding factor AML (CBF-AML; n=422), acute promyelo-

cytic leukemia (APL; n=312), KMT2A-rearranged AML (KMT2AR-AML; n=371), pAML (n=4502), sAML (n=832) and therapy-related AML (tAML; n=349). The patients' baseline, clinical/ treatment response and cytogenetic information are presented in **Table 2.1**. Mutational profiling identified 13,879 somatic mutations of variant allele frequency (VAF) $\geq$ 1% in the selected uniformed gene panel **Table A2,A3**. Using multivariate Cox modeling, we identified specific genomic lesion associations with survival. This approach enabled feature partitioning into "favorable vs. adverse" risks within diverse AML groups **Table A4**. Because the role of recurrent balanced translocations in AML diagnostics and the prognosis of tAML are already well-established, we focused our analyses on 5334 pAML + sAML cases without these pathognomonic lesions, hence, we excluded CBF-AML, APL, KMT2AR-AML, and tAML. Our objective was to determine if unique configurations of specific genetic lesions can produce distinguishable diagnostic patterns of pAML vs. sAML or within AML subsets including normal karyotype (NK-AML; n=3176) and abnormal karyotype AML (AK-AML; n=2158). This strategy was motivated by the observation of significantly different pAML vs. sAML survival (**Fig.** 2.1 A-C). Indeed, the supervised analyses yielded distinct clinical (**Table A5**) and genomic features that characterized each subtype (**Fig.** 2.1 D). Patterns detected by ULR and MLR included mutations in *CEBPA* (both monoallelic 'CEBPA^{Mo}' and biallelic 'CEBPA^{Bi}'), *DNMT3A, FLT3ITD, FLT3TKD, GATA2, IDH1, IDH2*^{R140}*, NRAS, NPM1* and *WT1* being enriched in pAML while mutations in *ASXL1, RUNX1, SF3B1, SRSF2, U2AF1*, -5/del(5q), -7/del(7q), -17/del(17P), del(20q), +8 and complex karyotype being prevalent in sAML (**Fig.** 2.1 D-F). Mutation burdens were similar in both AML subtypes (median: 2 mutations/individual; **Fig. A1**). The analyses of NK-pAML vs. NK-sAML (**Fig. A2**) and AK-pAML vs. AK-sAML (**Fig. A3**) revealed significant genetic associations that characterized each subset. In addition, clonal hierarchy analyses differentiated pAML vs. sAML based on founder and subclonal hits (**Fig.** 2.1 G). Despite these significant findings, the genomic profiles of pAML vs. sAML identified by MLR resulted in only 0.74 cross-validation accuracy of predictive performance when used

to reproduce pathomorphologic AML subtypes (**Fig.** 2.1 H).

2.3.2 Unsupervised genomic analysis unveils novel molecular AML groups spanning sAML/pAML dichotomy

As the accuracy of MLR prediction was modest, we explored other machine learning approaches as an alternative analytical strategy. BLCA of AML cases with complete mutational screens (**Table. A2**, n=2681) uncovered 4 novel genomic subgroups (**Fig.** 2.2A) based on the highest silhouette value (**Fig.** A4,A5). The biologic relevance of these subgroups was reflected in significantly different survivals [median (95% confidence interval)]: i) Genomic cluster-1 (GC-1) ; 34.1 (25.2-50.5) months], ii) GC-2; 26.5 (22.9-31.0) months], iii) GC-3; 15.8 (13.3-18.0), and GC-4; 9.2 (7.4-11.6) months (**Fig.** 2.2B,A6) and survival probabilities (**Table**. A6). Of note, the implementation of survival analyses was considered only to reflect on the biological and prognostic relevance of these clusters and not to replace current prognostic schemes. Moreover, the robustness of the BLCA clustering with respect to VAF was further validated when considering a higher cut-off of 15% which also resulted in 4 genomic clusters with a silhouette value of 0.86 and adjusted Rand index of 0.84 (**Fig**. A7).

2.3.3 pAML and sAML composition within genomic clusters

The distribution of genomic clusters within pAML and sAML was variable (**Fig.** 2.2C). For instance, pAML cases showed similar percentages of GC-1 (32%), GC-2 (33%) and GC-3 (25%) but fewer cases of GC-4 (10%). In contrast, sAML cases had higher percentages of GC-4 (22%) but lower GC-1 (5%) than pAML (**Fig.** 2.2C). The few GC-1 sAML cases may be suggestive of a possible subtype misclassification on presentation or an impact of an important genetic alteration (**Table.** A7). Higher percentages of patients with molecular good prognosis were found in pAML (GC-1/2; 65%) while sAML had more of higher risk cases (GC-3/4; 66%). Results of reverse analysis of normalized frequencies of pAML and sAML within cluster groups were consistent with the aforementioned results (**Fig.** 2.2D)

Table 2.1: **Baseline, clinical and cytogenetic characteristics of Acute Myeloid Leukemia cohort by subtype**

Variables	CBF-AML n (%)	APL n (%)	KMT2A-AML n (%)	pAML n (%)	sAML n (%)	tAML n (%)	All n (%)
Total population	422 (6,2)	312 (4.6)	371 (5.5)	4502 (66.3)	832 (12.2)	349 (5.2)	6788 (100)
Age (y) (median/range) *	52 (18-86)	51.8 (18-86)	63.8 (18-87)	66.9 (18-89)	70 (21-89)	67.3 (18-89)	66.2 (12-89)
$\geq 60y$	103 (32.9)	93 (34.9)	165 (60.8)	2374 (65.5)	638 (81.3)	183 (66.5)	3556 (64.2)
Gender							
Male	239 (56.6)	163 (52.2)	199 (53.6)	2373 (52.7)	535 (64.4)	142 (40.7)	3651 (53.7)
Female	183 (43.4)	149 (47.8)	172 (46.4)	2129 (47.3)	297 (35.6)	207 (59.3)	3137 (46.3)
Hematological and BM parameters*							
WBC ($10^9/L$) (median/range)	15.5 (0.1-351)	2.9 (0.3-155)	15 (0.4-427)	20.2 (0.1-600)	5.3 (0.5-388)	7.4 (0.5-303)	14.7 (0.1-600)
$\leq 3x10^9/L$	41 (10.4)	136 (50.7)	85 (22.9)	874 (20.4)	279 (36.6)	94 (30.9)	1509 (23.4)
Hemoglobin (g/dL) (median/range)	8.9 (2.5-19)	9.8 (2.7-16.4)	9.1 (3.5-18.5)	9.2 (2.3-17.9)	9.3 (5-16.5)	9.4 (3.4-16)	9.1 (2.3-19)
$\leq 10g/dL$	277 (69.8)	144 (52.7)	209 (66.1)	2479 (65.9)	484 (66.3)	188 (64.4)	3781 (65.6)
Platelets $10^9/L$ (median/range)	42 (3-529)	30 (1-228)	71 (2-578)	73 (2-2366)	50 (5-869)	53 (5-570)	54 (1-2366)
$\leq 10^{10}/L$	356 (83.6)	239 (87.5)	221 (61.4)	2663 (60.6)	573 (76.5)	222 (75)	4274 (75.4)
BM blasts % (median/range)	51 (20-99)	76 (20-100)	70 (20-100)	61 (20-100)	30 (20-97)	60 (20-99)	50 (20-100)
Antecedent diagnosis*							
MDS	–	–	–	–	627 (75.5)	–	627 (9.2)
MDS/MPN	–	–	–	–	58 (7)	–	58 (0.8)
MPN	–	–	–	–	39 (4.7)	–	39 (0.6)
ELN risk stratification							
Favorable (%)	422 (100)	312 (100)	0 (0)	656 (15)	25 (3)	23 (7)	1438 (21.2)
Intermediate (%)	0 (0)	0 (0)	0 (0)	2166 (48)	358 (43)	189 (54)	2713 (40)
Adverse (%)	0 (0)	0 (0)	371 (100)	1680 (37)	449 (54)	137 (39)	2637 (38.8)
Cytogenetics							
Normal	–	–	45 (12.1)	2812 (62.5)	364 (43.8)	111 (31.8)	3332 (49.1)
t(1;22)	0 (0)	0 (0)	0 (0)	3 (0.07)	0	0	3 (0.04)
inv(3)/t(3;3)	0 (0)	0 (0)	0 (0)	61 (1.4)	16 (1.9)	3 (0.9)	80 (1.2)
-5/del(5q)	5 (1.2)	2 (0.6)	54 (14.5)	288 (6.4)	130 (15.6)	40 (11.5)	519 (7.6)
t(6;9)	0 (0)	0 (0)	0 (0)	24 (0.5)	2 (0.2)	1 (0.2)	27 (0.04)
-6/del(6q)	0 (0)	2 (0.6)	4 (1.1)	21 (0.5)	12 (1.4)	9 (2.6)	48 (0.7)
-7/del(7q)	6 (1.4)	3 (1.0)	31 (8.3)	298 (6.6)	109 (13.1)	54 (15.5)	501 (7.4)
-9/del(9q)	33 (7.8)	4 (1.2)	8 (2.1)	82 (1.8)	10 (1.2)	6 (1.7)	143 (2.1)
del(12p)	1 (0.2)	4 (1.3)	16 (4.3)	113 (2.6)	18 (2.1)	19 (5.4)	171 (2.5)
del(13q)	2 (0.5)	2 (0.6)	6 (1.6)	35 (0.7)	15 (1.8)	8 (2.3)	68 (1.0)
del(16q)	1 (0.2)	1 (0.3)	2 (0.5)	20 (0.4)	4 (0.5)	0 (0)	28 (0.4)
-17/del(17p)	1 (0.2)	1 (0.3)	18 (4.8)	133 (3.0)	45 (5.4)	24 (6.9)	222 (3.3)
del(20q)	0 (0)	0 (0)	10 (2.7)	67 (1.5)	33 (3.9)	9 (2.6)	119 (1.8)
+8	35 (8.3)	44 (14.1)	82 (22.1)	405 (9)	117 (14.1)	40 (11.5)	723 (10.7)
-X	27 (6.4)	0 (0)	0 (0)	18 (0.4)	8 (1.0)	7 (2.0)	60 (0.9)
-Y	64 (15.2)	2 (0.6)	9 (2.4)	79 (1.7)	31 (3.7)	9 (2.6)	194 (2.9)
Complex	38 (9)	11 (3.5)	93 (25.1)	451 (10)	164 (19.7)	72 (20.6)	829 (12.2)

* Some data were not available

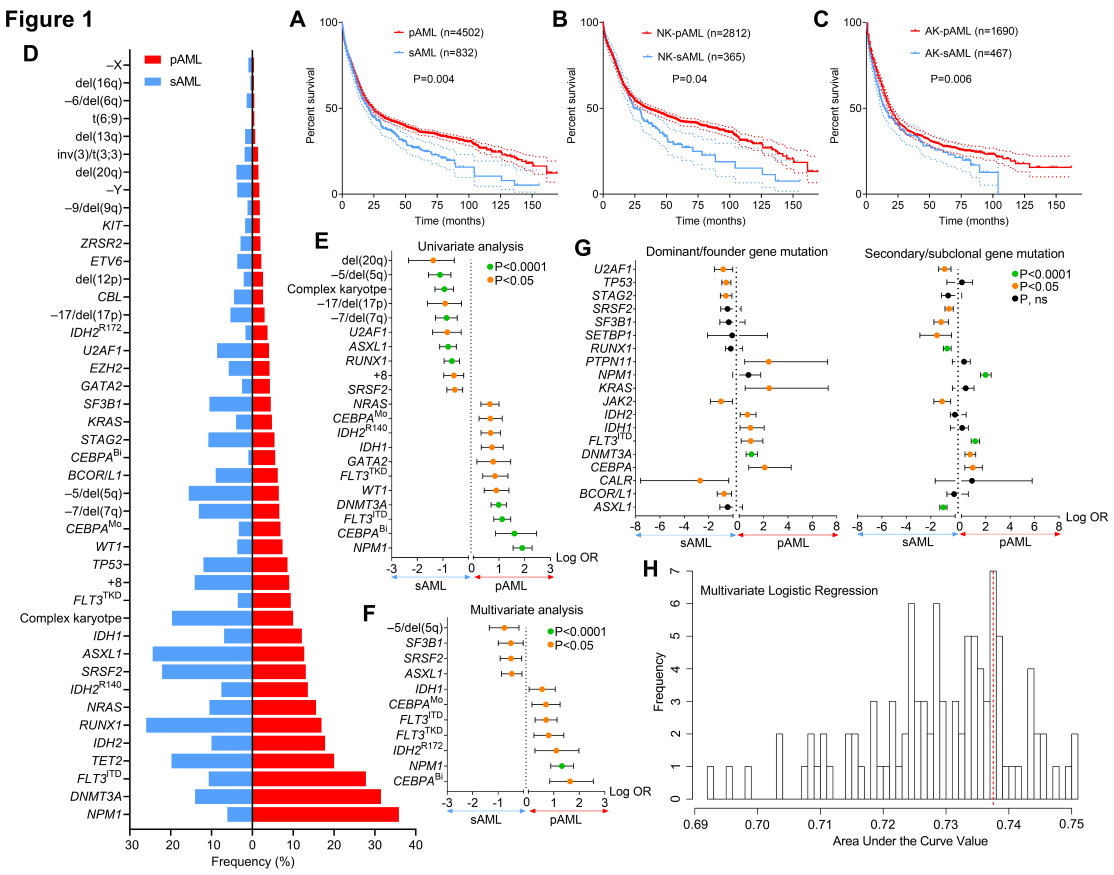


Figure 2.1: **Survival outcomes and mutational landscape of primary (pAML) versus secondary acute myeloid leukemia (sAML)**. (A-C) Kaplan-Meier survival curves of (A) pAML vs. sAML, (B) normal karyotype pAML (NK-pAML) vs. normal karyotype sAML (NK-sAML) and (C) abnormal karyotype pAML (AK-pAML) vs. abnormal karyotype sAML (AK-sAML). (D) A bar graph showing the frequency (in percent) of somatic mutations in pAML vs. sAML. (E) and (F) are forest plots representing univariate and multivariate logistic regression analyses showing the odds ratio (in log-scale) of the association of somatic mutations in pAML vs. sAML, respectively. (G) Forest plots representing univariate analyses showing the odds ratio (in log-scale) of the association of dominant/ancestral and secondary/subclonal somatic mutations in pAML vs. sAML, respectively. Levels of statistical significance, indicated by green, orange, and black (P<0.0001, P<0.05, and P>0.05, respectively), were obtained by Fisher's exact test. (H) Histogram of predictive performance (μ 0.74) of multivariate logistic regression using cross-validation area under the curve (AUC), i.e. we correctly predicted pAML and sAML classification in 74% of AML cases in our cohort using the distinct genomic characteristics of each subtype.

showing increased pAML proportion in GC-1 (89 vs. 11%) and sAML in GC-4 (67 vs. 33%). In addition, survival analyses within the same prognostic group showed no significant difference between pAML and sAML cases except in GC-4 group (**Fig.** A8; see **Fig.** A9 for

P-values of all pairwise comparisons of survivals of our 8-cluster x pAML/sAML groups).

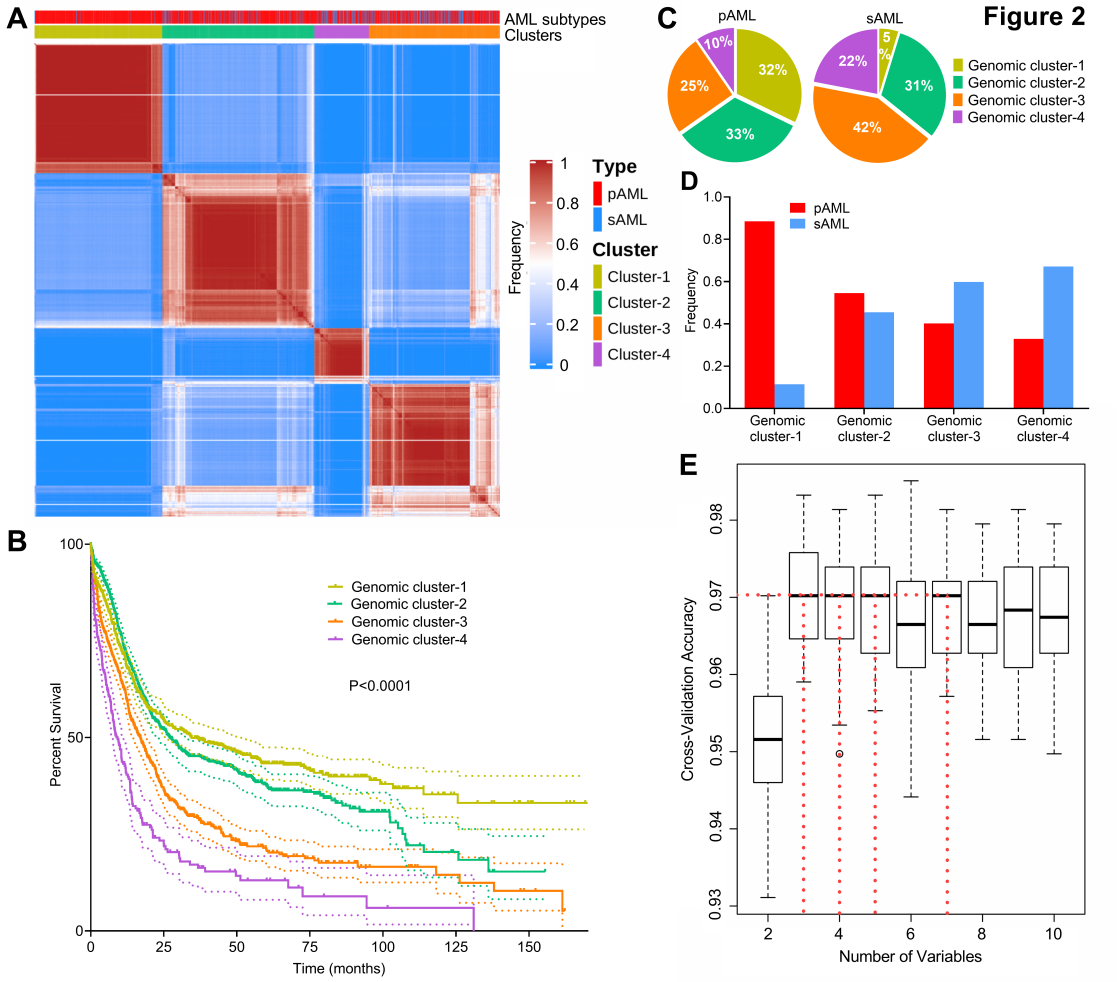


Figure 2.2: **Novel genomic clusters of acute myeloid leukemia (AML) identified by unsupervised analyses.** (A) Consensus matrix generated by applying latent class analysis on 1000 subsamples representing the frequency of two observations being clustered in the same group. (B) Kaplan-Meier analysis showing the overall survival (in months) of each genomic cluster (1-4). (C) A pie chart showing the percentage of cases belonging each genomic cluster (1-4) in primary (pAML, left pie) and secondary (sAML, right pie) AML. (D) A bar graph showing the frequencies of pAML and sAML patients in the genomic clusters after normalizing the samples by bootstrapping. (E) Hyperparameter selection plot for random forest modeling; cross-validation accuracy (CVA) is shown on the y-axis. CVA saturation in this plot indicates that 3 variables suffice to achieve the maximal accuracy of 0.97, i.e., this model correctly assigns 97% of AML cases prognosis in our cohort using their corresponding genomic features.

2.3.4 Invariant genomic features accurately predict molecular class assignments in AML

To link each cluster to its pathogenetic features, we generated an RF model. The resulting multiclass classifier which yielded a cross-validation accuracy of 0.97 (**Fig.** 2.2E). The model's globally most important genomic features, quantified by mean decrease in accuracy, included $NPM1^{MT}$, $RUNX1^{MT}$, $ASXL1^{MT}$, $SRSF2^{MT}$, $TP53^{MT}$, -5/del(5q), $DNMT3A^{MT}$, -17/del(17p), $BCOR/L1^{MT}$ and others (**Fig.** A10). Comprehensive group-specific observations showed that GC-1 was characterized by the highest prevalence of NK-AML (88%) and full presence of NPM1MT (100%; 86% with VAF>20%) that co-occurred with *DNMT3A* (52%), $FLT3^{ITD}$ (27%; 91% with VAF <50%), $IDH2^{R140}$ (16%, while absent $IDH2^{R172K}$) mutation with depletion or absence of *ASXL1, EZH2, RUNX1, TP53* mutations and complex cytogenetics (**Fig.** 2.3A). GC-2 had a higher percentage of AK-AML cases than GC-1, the highest frequency of CEBPABi (9%) and $IDH2^{R172K}$ (4%), $FLT3^{ITD}$ (14%) and $FLT3^{TKD}$ (6%) mutations occurring without $NPM1^{MT}$, while absent *ASXL1, RUNX1 and TP53* mutations (**Fig.** 2.3B). GC-3 had the highest frequency of *ASXL1* (39%), *BCOR/L1* (16%) and *DNMT3A* without *NPM1* (19%) mutation, in addition to being highly enriched with *EZH2* (9%),*RUNX1* (52%), *SF3B1* (7%), *SRSF2* (38%) and *U2AF1* (12%) mutations (**Fig.** 2.3]C). Of note, GC-3 showed a higher degree of heterogeneity. In fact, 53 cases in GC-3 had a silhouette value < 0 and of them, 15 cases were misclassified by the RF model. Further investigation of these misclassified cases showed that they had a wild type *RUNX1* mutation status while *RUNX1* mutation was prevalent in GC-3 (**Fig.** A11). Finally, GC-4 had the highest prevalence of AK-AML [96%; mostly of complex karyotype (76%)] and $TP53^{MT}$ (70%) that were associated with -5/del(5q) (68%), -7del(7q) (35%), -17del(17p) (31%) (**Fig.** 2.3D). Signature patterns, their importance and pairwise co-occurrences with other genomic markers, in addition to the clonal hierarchy of driver mutations in each cluster, are described in **Fig.** 2.3B-E,A12 and **Fig.** A13A-D, respectively.

We also analyzed the percentages of novel groups among each genomic lesion population

(**Fig.** A14,A15,A16). GC-1 represented 97% of *NPM1*, 50% of $FLT3^{ITD}$, 54% of *DNMT3A*, 43% of *IDH1*, and 43% of $IDH2^{R140}$ mutations as well as 43% of NK-AML; GC-2 accounted for 91% of $CEBPA^{Bi}$, 46% of *GATA2*, 50% of *WT1* mutations; GC-3 had 90% of *ASXL1*, 82% of *BCOR/BCORL1*, 52% of CBL, 53% of *ETV6*, and 46% of $IDH2^{R172K}$ mutations. It also represented the majority of splicing factor mutations (48% of *SF3B1*, 86% of *SRSF2*, 70% of *U2AF1*, and 65% of *ZRSR2* mutations), 98% of *RUNX1* mutations and the highest portion of del(20q) (65%) and trisomy 8 (49%); GC-4 represented 94% of *TP53* mutations, 62% of complex cytogenetics, 92% of -5/del(5q), 100% of -6/del(6q), 88% of del(12p), 91% of del(16q), and 92% of -17/del(17p).

When the clinical and baseline characteristics of each group were studied (**Table.** A8), GC-1/2 were found to contain a significantly younger age population compared to GC-3/4 (median age: 61 vs. 70 y, $p < 0.0001$, **Fig.** A17). Moreover, lower numbers of white blood cells correlated with higher risk disease ($p < 0.0001$, **Fig.** A18), possibly due to GC-3/4 harboring more dysplastic features than GC-1/2 groups, which had more proliferative AML phenotype.

Finally, we revisited the previously excluded well-defined prognostic AML groups and applied BLCA which demonstrated that APL, CBF-AML and t-AML constituted of a single genomic cluster each while 2 genomic groups were uncovered in KMT2AR-AML (**Fig.** A19A-B), including i) GC-A (median OS: 20.3 months) and ii) GC-B (median OS: 6.9 months) of distinct survival analysis (**Fig.** A19C). The most important genomic markers extracted by the RF model included *TP53* mutation followed by -5/del(5q), -7del(7q), -17del(17p) (**Fig.** A19D). The GC-2 KMT2AR-AML was characterized by the enriched presence of *TP53* mutation (79%), -5/del(5q) (63%), -7del(7q) (38%), +8 (38%) and -17del(17p) (31%) (**Fig.** A19E) while GC-1 KMT2AR-AML had absence/depletion of the aforementioned genomic aberrations except for +8 (18%), in addition to frequent *DNMT3A* (20%) and NRAS (16%) mutations (**Fig.** A19E).

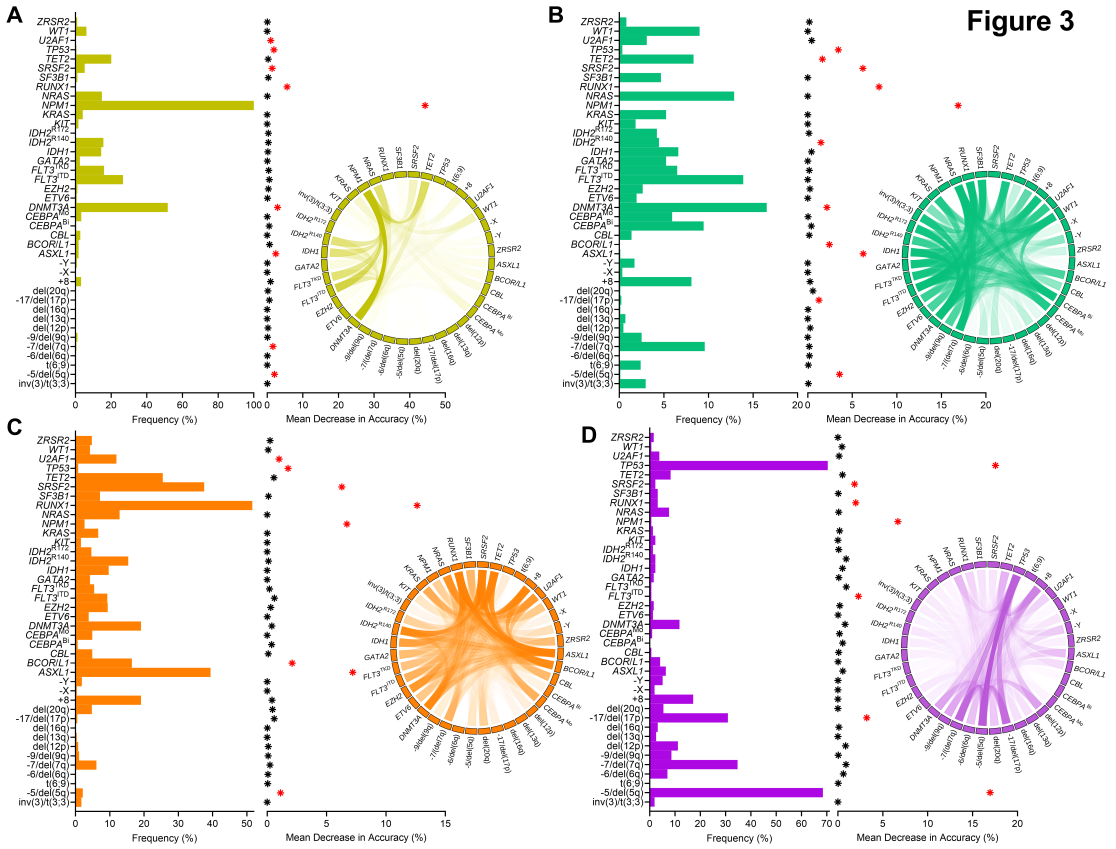


Figure 2.3: **Invariant genomic features driving each genomic group.** Bar plots representing the mutational profiles of (A) genomic cluster-1, (B) genomic cluster-2, (C) genomic cluster-3 and (D) genomic cluster-4 and their importance. Red asterisks represent the most important genomic features based on an arbitrary importance cutoff of $\geq$ 0.01 0.01 mean decrease in accuracy. In addition, circos diagrams showing the pairwise co-occurrence of mutations in each genomic cluster are illustrated to the right of the bar graphs. The color code of circos diagrams correspond to the genomic clusters. The percentage of a co-occurrence between the first and the second gene mutations is represented by the color intensity of the ribbon connecting both genes.

2.3.5 Automated cluster predictor and confirmatory studies

We performed internal and external validation of our genomic clustering model. The internal confirmatory cohorts consisted of randomly selected training (80%, n=2144) and test (20%, n=537) sets. BLCA and RF were applied on the training set only and the survival analysis of test set was separately done (**Fig.** A20). The test cohort did not show significant survival differences per each genomic cluster as compared to the training set when Kaplan-Meier analyses were performed **Fig.** A21A-D). We further evaluated how

the number of the identified clusters varies across subsets of training cohort by randomly sampling 75% of observations to prevent overfitting. Silhouette values across random samples showed 4 as the optimal number of clusters (**Fig.** A22). External validation was then conducted using an independent cohort of 203 AML patients from the MDACC (pAML, n=143; sAML, n=60) with a median follow up of 12 months (0.1 - 35.3 months) and fully annotated characteristics (**Table.** A9). Gene sequencing of the selected gene panel identified a total of 723 somatic mutations in the MDACC cases (**Table.** A10). The Kaplan-Meier survival analyses of the original data and MDACC cases showed similar survival among each genomic cluster assigned by the RF model (**Fig.** 2.4A-D). Details of further validation approaches for hyperparameter tuning and depth selection for random forest model is provided in the supplementary section. Furthermore, our genomic subclassification model is available as a web-based open-source resource that can be accessed widely by clinicians and the public to forecast the subclassification and estimated survival of AML patients without known pathognomonic lesions, balanced-translocations or tAML (`https://drmz.shinyapps.io/local_app/`) (**Fig.** 2.4E). A conceptual framework summarizing our overall approach is illustrated in **Fig.** A23.

2.4 Discussion

Apart from certain well-defined AML subtypes (e.g. CBF-AML, APL, and KMT2AR-AML), historically, AML patients have been subcategorized into subgroups defined by pathomorphological features and broad anamnestic clinical criteria due to the inability to precisely infer the presence of antecedent prodromal disease [52, 64]. Ubiquitous application of genomic diagnostics has provided opportunities for objective sub-classification of AML, which due to its mechanistic foundation can direct discovery and application of molecularly targeted therapeutics and allows for tailored personalized risk-stratification [4]. Building on this potential and the power of modern genomic and bioanalytical approaches, we investigated whether rational genomic tools would yield precise, simple, and diagnostic

Figure 4

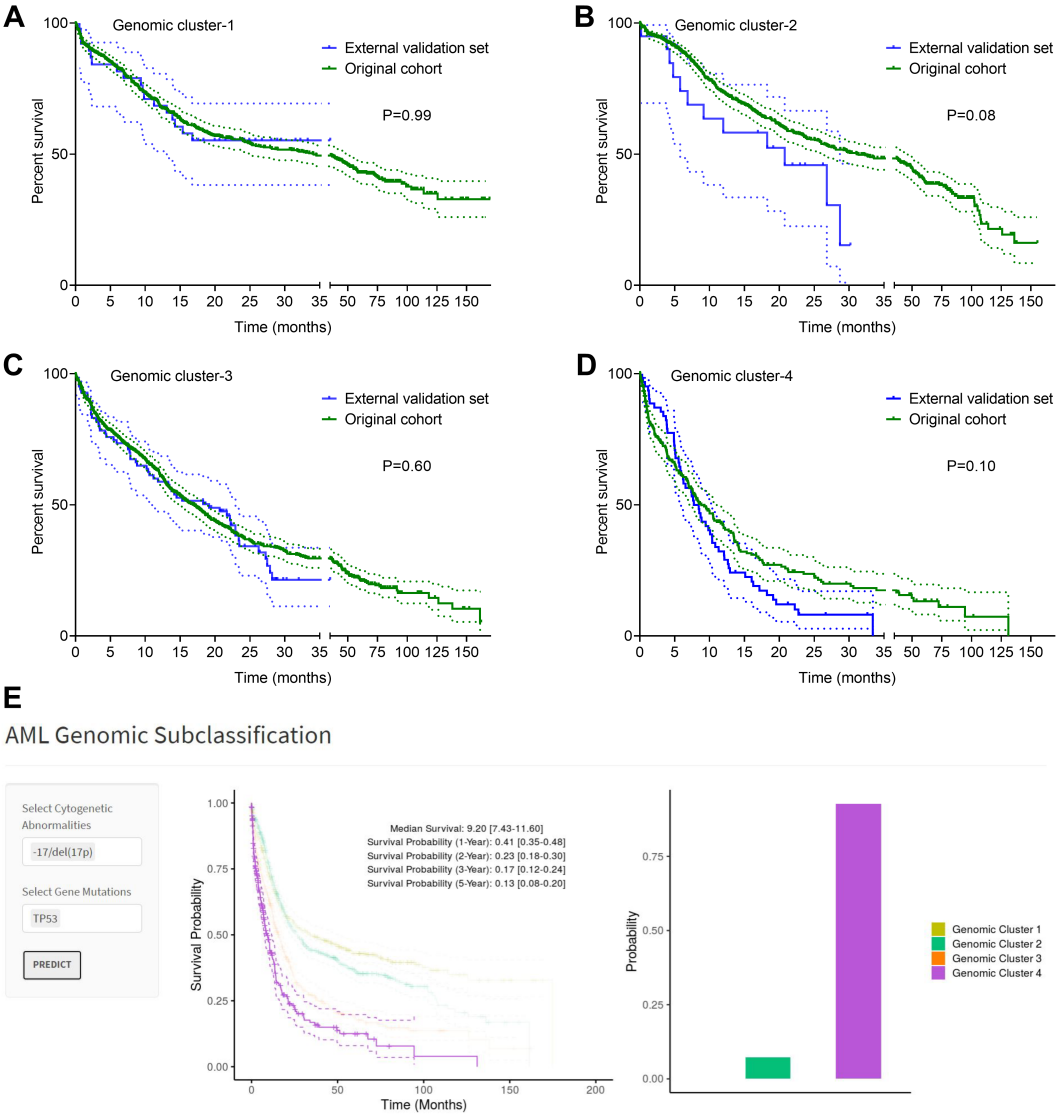


Figure 2.4: **Model validation and uniform resource locator.** (A-D) Kaplan-Meier survival analyses (time in months) for the external validation of the model using external data from the MD Anderson Cancer Center (MDACC) vs. the original data, is represented in each cluster: (A) genomic cluster-1, (B) genomic cluster-2, (C) genomic cluster-3, and (D) genomic cluster-4. (E) A screenshot of the website interface to our model.

AML subclasses that would reflect genomic pathogenesis and prognosis. This is best illustrated by the ability of genomic clusters to redefine historical subclasses such as pAML and sAML, which have been shown to only variably correlate with genetic patterns and pathogenesis. By applying machine learning methods to an unprecedented large cohort

of AML patients with detailed molecular annotations, we developed a new classification model. Credence in this model is fortified by its accurate classification (97% correct) and the plausibility of the distinctive genomic features that contributed to the overall assignment accuracy. Utility of this model is supported by different cluster group survival outcomes. Feature association with both pathobiology (molecular parameters) and survival suggest that this model could overcome some limitations of previous pathomorphological based classifications of pAML and sAML.

Our results reaffirm previous studies and extend them by integrating more molecular features and expanding diagnostic and pathogenetic implications [10, 49, 52, 56]. In particular, we focused on the inclusion of genomic signatures, despite their variable degree of global importance to achieve the highest possible genomic classification accuracy. For instance, and in line with previous reports,10,13,25-28 NPM1 and TP53 mutations contributed greatly to creating the lower and higher risk phenotypes (corresponding to the clinical/survival risk), respectively [4, 54, 65–68]. However, the highest cumulative accuracy was only achieved by the incorporation of the status (+/-) of additional genomic lesions including *RUNX1, ASXL1, SRSF2,* and *DNMT3A* mutations, -5/del(5q), -17/del(17p), and others. Our molecular 4-tiered model is not meant to challenge or replace previously established prognostic schemes. It mainly focuses on objectively subclassifying genomic-undefined pathomorphological AML subtypes including pAML and sAML. However, as we compare it to traditional prognostication tools that incorporate prognostic genomic features, like the 2017 European LeukemiaNet (ELN-2017), we would like to point out certain advantages that were concluded from our genomic cluster-based model. Our model expands a larger pool of genomic signatures and quantifies their corresponding importance. The latter is crucial when determining the probability of objective subclassification in complex heterogeneous AML cases harboring combined ELN-2017 defined favorable and adverse genomic lesions. Strikingly, the model describes distinct clustering of a variety of previously described genomic lesions that are known to influence AML outcome and emerged

as AML cluster determinants. *DNMT3A, IDH*^{R140}, and *TET2* mutations, only when occurring without *NPM1*^{MT}, are important genomic determinants of GC-2. Splicing factor mutations (*SRSF2, U2AF1*) contributed substantially to our model's performance and were noticeably enriched in the GC-3, indicating their predicted potential to be a distinct AML subgroup [52, 62]. Moreover, *RUNX1*^{MT} had the second-highest global importance, and crucially contributed to the identification of all novel groups. Specifically, *RUNX1*^{MT} was highly prevalent in the GC-3. Consequently, our data confirm the substantial presence of *RUNX1*-mutant AML in the most recent WHO classification as a provisional disease category [13, 69]. Interestingly, *BCOR/L1* mutations emerged as a potential genomic marker of GC-3. Although CK-AML was abundant in the poorest survival group like defined in ELN-2017, the concurrent presence of other important genomic markers identified by our model [-5/del(5q), -17/del(17p) and *TP53*^{MT}], seemed to delineate its classification. When these aforementioned markers were absent, CK-AML was also seen in other groups (GC-3). Hence, our model argues that genomic subclassification of CK-AML is strongly dependent on the present/absent status of other decisive correlating genomic markers. Finally, the model is dynamic and displays flexibility and personalization by accounting for accurate probabilities of assignment to each cluster per the presence/absence of each genomic feature and its interactions with other signatures, rather than predicting a single classification. It also defines the estimated survival interval of each genomic group, which can be considered when assessing prospective AML patients' prognoses. Due to the mechanistic focus, the deliberate exclusion of certain clinical data may appear as a limitation to our model. While we believe that some of the phenotypic features are a result of the genomic makeup and are likely codependent, we acknowledge that selected parameters may be later incorporated similar to the genomic features to be discovered in the future. The latter may, for example, include some of the germ-line alterations, clonal/subclonal burden, or configuration of hits as demonstrated for *CEBPA*^{Bi} mutations. Furthermore, some genomic clusters showed a higher degree of heterogeneity as compared to other, which can be likely improved by future

incorporation of more complex models such as neural network-based clustering or the use of infinite priors in Bayesian setting where a larger cohort is available. Besides, as all of the patients receive therapy, new effective drugs could affect prognosis and thus may have a global subgroup-specific impact on survival and the predictive value of survival curves within subgroups may have a limited shelf life. Although the predictive accuracy of our genomic model was validated and our approach accounted for possible generalizability limitation by including multicenter cohorts, eventually prospective external validations of longer patients' follow-up durations are still warranted. Additionally, we envision that molecularly based risk assessment may have rational implication on the use of specific therapy choices especially when targeted agents and their combinations will be more widely applied and thus purely clinical classification schemes will become obsolete to provide generalizable survival predictions. In conclusion, our study demonstrates that despite the tremendous heterogeneity of AML genomics, non-random genomic relationships, captured by machine learning methods, are capable to accurately assign objective molecular classification and prognosis irrespective of the availability of clinicopathologic or anamnestic information. It clearly indicates that classical distinction of sAML from pAML cannot be justified on molecular levels and rather molecular signatures/patterns should provide a prevailing impetus for classification schemes. Our model provides a personalized genomic tool for AML subclassification that is publicly shared.

Chapter 3

MOLECULAR PATTERNS IDENTIFY DISTINCT SUBCLASSES OF MYELOID NEOPLASIA

Genomic mutations drive the pathogenesis of myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML). While morphological and clinical features, complemented by cytogenetics, have dominated the classical criteria for diagnosis and classification, incorporation of molecular mutational data can illuminate functional pathobiology. We combined cytogenetic and molecular features from a multicenter cohort of 3588 MDS and secondary AML patients to generate a molecular-based scheme using machine learning methods and then externally validated the model on 412 patients. Molecular signatures driving each cluster were identified and used for genomic subclassification. Unsupervised analyses identified 14 distinctive and clinically heterogenous molecular clusters (MCs) with unique pathobiological associations, treatment responses, and prognosis. Normal karyotype (NK) was enriched in MC2, MC4, MC6, MC9, MC10, and MC12 with different distributions of *TET2, SF3B1, ASXL1, DNMT3A*, and *RAS* mutations. Complex karyotype and trisomy 8 were enriched in MC13 and MC1, respectively. We then identified five risk groups to reflect the biological differences between clusters. Our clustering model was able to highlight the significant survival differences among patients assigned to the similar IPSS-R risk group but with heterogenous molecular configurations. Different response rates to hypomethylating agents (e.g., MC9 and MC13 [OR: 2.2 and 0.6, respectively]) reflected the biological differences between the clusters. Interestingly, our clusters continued to show survival differences regardless of the bone marrow blast percentage. Despite the complexity of the molecular alterations in myeloid neoplasia, our model recognized functional objective clusters, irrespective of anamnestic clinico-morphological features, that reflected disease evolution and informed classification, prognostication, and molecular interactions.

3.1 Introduction

The myelodysplastic syndromes (MDS) are a collection of diseases encompassing a pathologically distinct, broad spectrum of myeloid disorders, some of which represent stages of the natural history of leukemia[13, 70]. Until now, morphological features, later enhanced by cytogenetic abnormalities, have dominated the pathology criteria for MDS diagnoses. These can be limited by inter-observer variability, restricted resolution, and lack of functional correspondence to molecular underpinnings[71, 72]. Widely-used MDS prognostic classification schemes may be convergent, as they group cases with similar features yet different molecular origins; or divergent, as they assign cases with similar molecular lesions into different pathomorphological sub-entities[6]. Moreover, when considering molecular features, morphology-based classifications overemphasize specific parameters (e.g., blasts), which may represent essentially the stage of the disease, as opposed to molecular evolution. As a result, blast-defined MDS subtypes would contain a mixture of cases with various molecular derivations[6, 73–76]. Classification schemes according to clinical features are more practical, but apart from the weight of cytogenetics on prognosis, clinical prognostication does not reflect the disease pathogenesis. The advent of next generation sequencing (NGS) has led to the discovery of a multitude of mutations in various genes, and recognition of the tremendous molecular diversity and clonal hierarchy within myeloid malignancies[57, 77, 78]. These factors, along with cytogenetics, constitute the underpinnings of MDS pathogenesis. Given their complexity, attempts to consolidate mutational patterns into broader disease sub-entities have been made, with conventional integrative approaches of classical, clinical, and pathomorphological features used as a gold standard in supervised analytic strategies. Consequently, the patterns of molecular features have been analyzed to fit into morphologic groups, with limited success given the complexity of mutations and their interactions, particularly with respect to disease progression[8, 9]. Therefore, new strategies may be needed to deconvolute this molecular diversity and generate subdivisions of patients with MDS whose disease fits within molecular pattern similarities,

better reflecting prognosis and which could then be targeted with specialized therapeutic approaches. Machine learning (ML) analytic methods, as demonstrated in acute myeloid leukemia (AML)[79], provide new opportunities to integrate the molecular pathogenesis by identifying relevant patterns, which could serve as molecular sub-entities[8, 9, 80, 81]. Here, we took advantage of a large, well-annotated cohort of patients with MDS and secondary AML (sAML) to test the hypothesis that related molecular patterns can be analyzed in an unbiased/unsupervised fashion to characterize molecularly-defined configurations of MDS/sAML. We used a similarity-based ML approach to cluster patients into molecular subgroups, further validated based on clinical features.

3.2 Methods

3.2.1 Patient Cohort

We assembled a large cohort of patients diagnosed with MDS and sAML to generate a comprehensive genomic data set. Patient data from the Cleveland Clinic ([CC], n=1627), The Munich Leukemia Laboratory ([MLL], n=1275), and publicly available data sets (The BEAT AML master trial and The EuroMDS cohort Patients, n=686)[8, 56] were combined to form a cohort of 3588 MDS and sAML patients (Supplementary Table 1). Targeted NGS results at the time of diagnosis were collected and adjusted to analyze the most common somatic myeloid mutations (Supplementary Table 2). Electronic medical records were reviewed to collect clinical parameters at the time of diagnosis and from resources accessible online from the publicly shared data sources (EuroMDS). Samples were collected after obtaining written informed consent according to the protocols approved by the respective institutional review boards (see Supplementary Materials).

3.2.2 Genetic Studies

For the data collected at CC, whole exome sequencing (WES) was performed as previously described[59, 60, 62]. Paired tumor and germline DNAs were used for WES. Data were validated using a TruSeq Custom Amplicon Kit (Illumina) (Supplementary Table 2). Vari-

ants were annotated using Annovar and filtered using an in-house bioanalytic pipeline[57–59, 62]. The gene sequencing methods of publicly shared MDS and sAML patients were previously described[8, 56]. For validation, an independent cohort of MDS/sAML patients (UT Southwestern medical center and Karmanos Cancer Institute was used; see **Table** B1 & Supplementary Methods).

3.2.3 Statistical Analyses

Our ML strategy was based on a consensus-clustering approach via autoencoders coupled with gaussian-mixture modeling (GMM)[82]. The resultant model was validated internally and externally on an independent cohort (detailed description in the Supplementary Materials). Our genomic subclassification model is available via web-based open-access resource as well (`https://drmz.shinyapps.io/mds_latent`).

3.3 Results

3.3.1 Unsupervised clustering of the molecular architecture of MDS and sAML reveals novel molecular subgroups regardless of histological or clinical features

Among the 3588 patients included in this cohort, 735 (20%) had sAML, 774 (22%) had higher-risk MDS (HR-MDS), and 2079 (58%) had lower-risk MDS (LR-MDS). Abnormal karyotype was found in 1548 cases (43%) (Table 1), and 2763 patients (77%) had at least one somatic mutation, with 284 cases (8%) harboring > 4 mutations (**Figure** B1). Using unsupervised ML analysis of the mutational panel in our cohort, we identified 14 molecular clusters (MC1-MC14) according to distinct genomic features (**Fig.3.1A,B**). The number of MCs was determined based on the highest silhouette value (**Fig.3.1A**). The MCs size varied; for example, 26% of the cases were assigned to MC2 and only 2% to MC3 (**Fig.3.1C**). The most distinctive clinical and molecular features defining the MCs were identified (Table 2, **Fig**.B2). Overall, the most important genomic features used in the model were quantified based on the mean decrease in accuracy (**Fig.**B3,B4). The resultant MC signatures are illustrated in **Fig.3.2**.

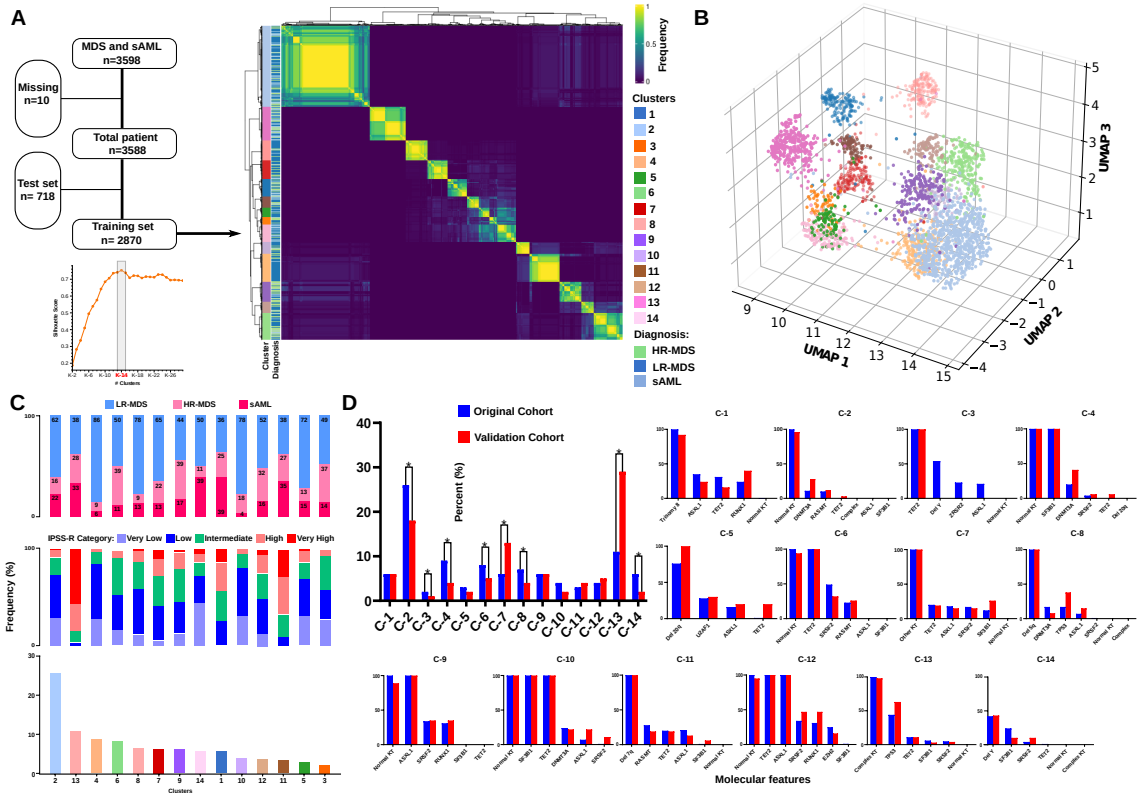


Figure 3.1: **Genomic clusters of myelodysplastic syndrome and secondary acute myeloid leukemia identified by unsupervised analysis.** **A)** Unsupervised clustering of binary mutation profiles performed through iteratively clustering sub-samples of original data and keeping track of the frequency of pairwise co-occurrence of samples in the same cluster. **B)** To visualize the clusters on a three-dimensional space, we have generated an exemplary dimension reducing space using UMAP coupled with the autoencoder model. A 16-dimensional linear embedding of binary mutation profiles was generated and reduced to 3d using UMAP in a nonlinear fashion. A specific figure legend color presents each genomic cluster. **C)** Bar graph showing the frequency of each cluster in our cohort (lower panel) and the relative frequency of low-risk myelodysplastic syndrome (LR-MDS), high-risk myelodysplastic syndrome (HR-MDS), and secondary acute myeloid leukemia (sAML), upper panel. The middle panel is showing the relative frequency of different Revised International Prognostic Scoring System (IPSS-R) among different clusters. **D)** Bar graph illustrating the frequency of each genomic cluster in the original and the validation cohort. Significant differences are indicated by asterisks. Graphs from C1-C14 illustrate the frequency of the most important molecular features in the original and the validation cohorts.

Our ML model performance was then validated internally and externally. For the internal validation, we randomly selected training (80%, n=2870) and test (20%, n=718) sets for K-fold cross-validation to assess the fit of our model and divided the cohort into five

folds. Based on the highest silhouette value in each fold, a total number of 14 clusters was optimal in all five folds. Adjusted-Rand Index (ARI) comparisons between the folds showed a minimum ARI of 0.85 (**Fig.**B5). The external validation was conducted using an independent cohort of 412 MDS/sAML patients (**Table.**B3) with a different patient clinical composition distinct from the original cohort. Based on the mean decrease in accuracy, we selected and compared the most important characteristics between the original and the validation cohorts. As expected, no significant differences in cytogenetics and molecular features in most MCs were observed (**Fig.3.1C**).

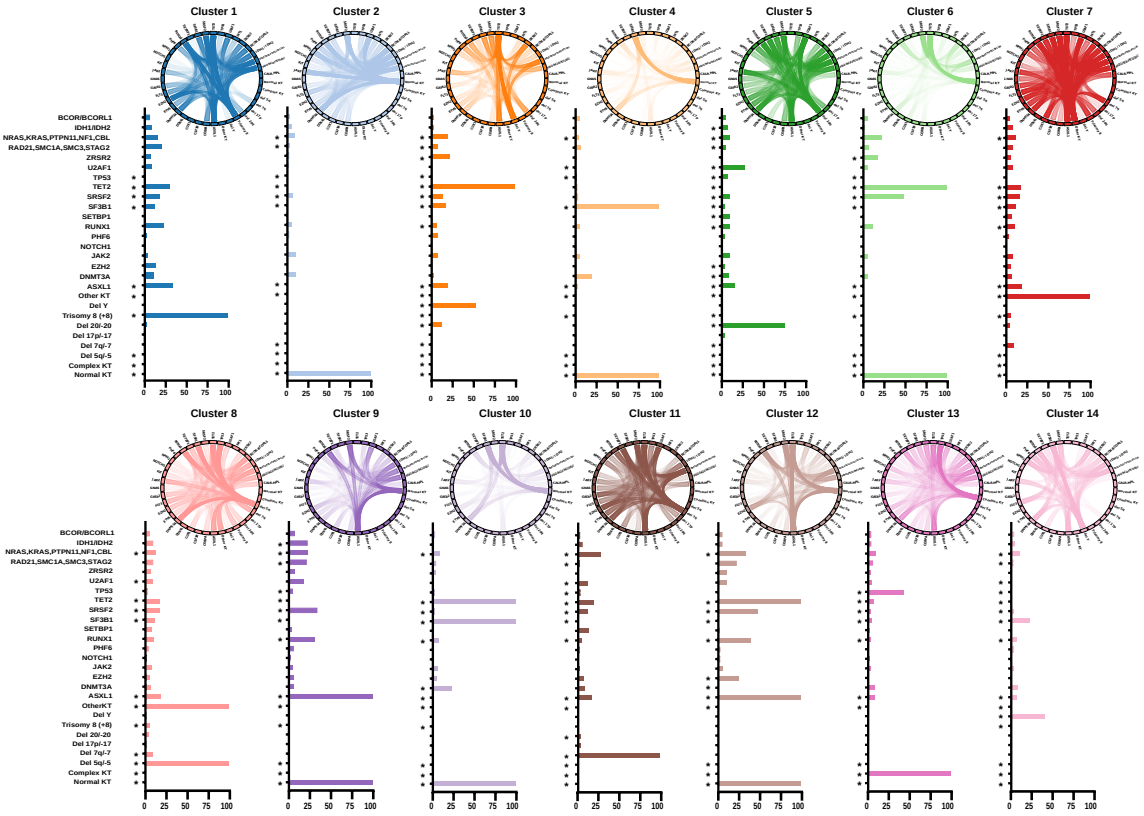


Figure 3.2: **Genomic features drive each genomic group.** Bar plots represent the mutational profiles of all genomic clusters (clusters 1 to 14) and their importance. Asterisks denote the most important genomic features based on an importance cutoff of a mean decrease in accuracy ≥ 0.01. The circos diagrams above each cluster show the pairwise co-occurrence of mutations in all clusters and are illustrated to the right of the bar graphs. The colors of circos diagrams correspond to the clusters. The percentage of mutational co-occurrence between first and second gene mutations is represented by the color intensity of the ribbon connecting both genes.

3.3.2 Molecular clusters composition

The composition of the MCs were distinct, reflecting differing morphological diagnoses and bone marrow (BM) blast counts (Table 2). For instance, LR-MDS patients comprised most of MC8 (78%), MC10 (78%), and MC5 (72%). In addition, more than 50% of the cases within MC2, MC4, MC6, MC7, and MC14 were LR-MDS patients. Conversely, HR-MDS and sAML cases comprised more than 30% of MC3, MC6, MC9, and MC12. When we applied reverse analysis, the majority of the sAML cases populated MC2 (28%), MC13 (18%), MC1 (11%), and MC14 (10%). HR-MDS cases were mainly classified in MC2 (19%), MC6 (15%), MC13 (14%), and MC9 (11%). Finally, LR-MDS clustered in MC2 (27%) and MC4 (13%; Supplementary Figure 6). Moreover, the distribution of different revised international prognostic scoring system (IPSS-R) risk groups among our MCs were distinct and heterogenous (**Fig.3.1C**, middle panel). Although most of the cases included in MC11 and MC13 were very-high and high-risk groups according to IPSS-R, both clusters continue to contain patients from other risk groups who share the same molecular configuration. Blast percentages in MCs were consistent with the risk distribution of cases, and the median blast percentage was consistent with the composition of each MC (Table 2 and **Fig.**B7). For instance, while MC1 and MC13 had a median blast percentage of $> 10\%$, MC2 and MC4 had a median of $< 5\%$, consistent with the enrichment of early-stage (LR-MDS) cases within the latter MCs.

3.3.3 Machine learning-derived clusters reflect functional relationships

Broad cluster-specific analyses revealed that all MC4 cases had NK and **SF3B1** mutations. Similarly, all MC10 cases had NK and *SF3B1* mutations in addition to *TET2* mutations (100%). *DNMT3A* mutations were present in 20% and 24% of MC4 and MC10, respectively. MC2, MC6, and MC8 demonstrated distinct genomic signatures: MC2 included cases with NK only (100%) and some *DNMT3A* (11%), *JAK2* (11%), and *RAS* pathway (10%) mutations; MC6 cases had similar features to MC2 but were also enriched in *SRSF2* (49%)

and *RAS* mutations (23%); MC8 was characterized by the presence of del5q/-5 (100%), *DNMT3A* (17%), and *TP53* (17%) mutations. MC3 included cases with *TET2* (100%), *ZRSR2* (23%), and *ASXL1* (21%) mutations with delY (54%). MC14 included cases with delY (42%) but without *TET2* mutations, distinct from MC3. In contrast, MC12 included cases with *TET2* (100%), *ASXL1* (100%), *SRSF2* (48%), *RUNX1* mutations (40%), and NK (100%) similar to MC9, which contained *ASXL1* (100%) with *SRSF2* (34%) and *RUNX1* (31%), but lacked *TET2* mutations. MC5 grouped cases with del20q/-20 (76%) and *U2AF1* mutations (28%). MC7 was characterized by other cytogenetic abnormalities, not including del5q/-5 compared to MC8. MC1 was characterized by trisomy 8 (100%), *ASXL1* (35%), *TET2* (31%), and *RUNX1* (24%) mutations. MC11 included cases with del7q/-7 (100%) and *RAS* pathway mutations (28%). Finally, MC13 contained cases with complex karyotype (100%) and *TP53* (44%) mutations. To understand the frequency of each mutation within the novel identified clusters, we also analyzed the distribution of each genomic mutation and cytogenetic abnormalities across clusters (**Fig.**B8).

3.3.4 MDS molecular clusters have clinical correlates

We explored differences in overall survival (OS) across the identified MCs (**Fig.3.3A,B**). As expected, the high degree of molecular heterogeneity translated to divergent survival in each group (**Fig.B9**). By grouping MCs according to survival impact, we distinguished 5 risk categories (**Fig.3.3C** and **Table.**B4), which were recapitulated in the external validation cohort (**Fig.3.3D**). In addition to survival, MCs also demonstrated distinct clinical differences. For instance, patients in MC1, MC11, and MC13 had significantly lower platelet counts (87, 48, and 76, respectively, p-value <0.001) compared to other clusters. We also detected discrete survival differences in patients (N=2863) treated with hypomethylating agents (HMAs) and/or allogeneic hematopoietic stem cell transplant. Even after accounting for the different treatments, the risk groups continued to show significant differences in OS (**Fig.**B10). Interestingly, we noticed interaction effects between treatments response and our

MCs. For instance, a higher response rate to HMAs (according to the International Working Group criteria24) occurred in patients assigned to MC9, MC10, and MC12 (36%, 33%, and 32%, respectively) compared to response rates in patients assigned to MC1, MC13, MC3, and MC7 (13%, 13%, 14%, and 15%, respectively; **Fig.3.3E,F**). Logistic regression analysis showed that MC9 (odd ratio [OR]: 2.2, 95%CI: 1.2-3.9) and MC13 (OR: 0.6, 95%CI: 0.4-0.9) were associated with different HMAs response rates.

The blast percentage within MCs did not appear to affect survival after 25 months. For instance, although MC13 contained 38% and 33% of LR and HR-MDS patients, respectively, the prognosis was homogenously worse when compared to other MCs. Using Cox-Proportional Hazard model accounting for relevant clinical variables, the assigned risk groups based on our clustering method showed significant survival differences (**Fig.**B10B). Compared to the Low-Risk group (OS [95% CI]; 93 months [42-132]), patients classified as Very High-Risk (OS [95% CI]; 9 [4-24]) High-Risk (OS [95% CI]; 17 [5-53]), Intermediate-High risk (OS [95% CI] 33 [12-92]), and Intermediate-Low risk (OS [95% CI]; 62 [19-188]) had significantly worse OS. Our clustering model was able to highlight the significant survival differences among patients assigned to the similar IPSS-R risk group but to different MCs (**Fig.**B11). For instance, we observed significant differences in OS among patients assigned to very low risk IPSS-R based on our MCs (HR:1.9, 95%CI: 1.5-2.8).

3.4 Discussion

While MDS classification schemes evolved as useful clinical diagnostic or prognostic tools, diagnostic criteria according to genomic signatures reflective of molecular pathogenesis have not been established[8, 13, 83]. Furthermore, previous attempts to incorporate mutations into prognostic schemes to increase their predictive precision resulted in considering only a handful of consequential mutations[79]. One of the reasons for the inability to establish reproducible genotype/phenotype associations might be the application of primarily supervised strategies using traditional statistics and clinical classifications (reliant

on subjective nosology and time-dependent parameters) as a gold standard. Indeed, the tremendous diversity and complexity of molecular lesions hamper the application of conventional bioanalytic methods. To overcome the limitations of these traditional approaches, our study applied modern ML strategies to objectively integrate molecular features able to decipher patient sub-cohorts with known and/or previously cryptic associations. This strategy was not meant to compete with or replace current well-established prognostication tools[6], but rather illuminate the genetic sub-classification of MDS and related conditions in an operator-independent fashion according to molecular correlations and mutual functional proximity. Despite the exclusion of anamnestic clinical criteria, the resultant scheme yielded a reproducible and validated system of genetically-related disease clusters reflective of the genomic pathogenesis and prognosis, irrespective of established standards. Notably, our molecular classification has enabled the recognition of cases with convergent molecular mechanisms, e.g., for the rational selection of suitable therapies. Moreover, the personalized risk stratification method is independent of disease duration and stage. It does not involve blast count, whose predictive weight dominates most of the older disease schemes and the recently proposed ML-based prognostication model[8, 13]. Our ML-based molecular model defines unique clusters according to the previously described genomic features and their combinations known to influence MDS and sAML outcomes[1, 8, 70, 84, 85]. Moreover, the analysis of the invariant cluster-defining molecular combinations points towards previously unsuspected relationships or convergent pathways. Illustrative examples of such molecular associations are presented in the supplementary materials (Supplementary Results). Unlike previous prognostication systems highly dependent on the blast count[6], our MCs were heterogeneous in this regard. This observation raises many questions about the implication of BM blast percentages on molecularly-based diagnoses. Indeed, our ML-based scheme indicates that BM blast may correlate more with the stage of the disease rather than the molecular architecture. For instance, although MC13 included patients with the worst prognosis, almost 1/3 of the cases in this cluster had low blast counts while sharing

a similar molecular makeup with sAML, reflecting different stages of the same disease. Analogous observations were made in other clusters containing molecularly similar patients at various points of their clinical course. Significant survival associations with BM blasts and MCs also suggest that these variables capture different information regarding the disease pathogenesis. It is important to emphasize that the recent attempts to integrate cytomolecular features into MDS classification for personalized approaches were also based on traditional clinical parameters, which always outweighed the variables derived from the genomic makeup. For instance, when analyzing the fraction of explained variation attributable to different prognostic factors for OS, BM blast percentage, age, and sex alone accounted for more than 50%. In comparison, molecular features only had limited power in the proposed model ($\tilde{3}0\%$)[8]. In our model, focusing on the objective molecular signature to characterize the features of different clusters with the exclusion of morphological and clinical data may seem a limitation. However, we believe that clinical and morphological features constitute the results of genetic hits. We showed that our molecular clustering of MDS successfully identified unique patterns of genomic associations and possibly uniform/similar pathogenesis even if individual connections cannot be rationally discerned on this junction. We acknowledge that additional parameters such as allelic configuration/burden, mutation types, clonal/subclonal burden, and germline predisposition may add a significant value if incorporated, perhaps helping to further sub-stratify some of the more heterogeneous clusters. Another limitation of any analytic strategy (supervised/unsupervised) is that less common mutations remain unappreciated because of the lack of statistical power. This is also a flaw of our approach, which we attempted to mitigate by combining mutations affecting the same functional pathways and identifying rare hits confined strictly to one cluster to allow for inferences in terms of their functional associations. In conclusion, despite the complexity and the diversity of molecular alterations in MDS and sAML, by deploying artificial intelligence, we were able to recognize functional and pathologically related, objective clusters irrespective of the anamnestic clinico-morphological features.

Our model provides molecular correlation for a better understanding of the pathobiological mechanisms of disease, progression to higher stages, and identification of future targets for novel therapies.

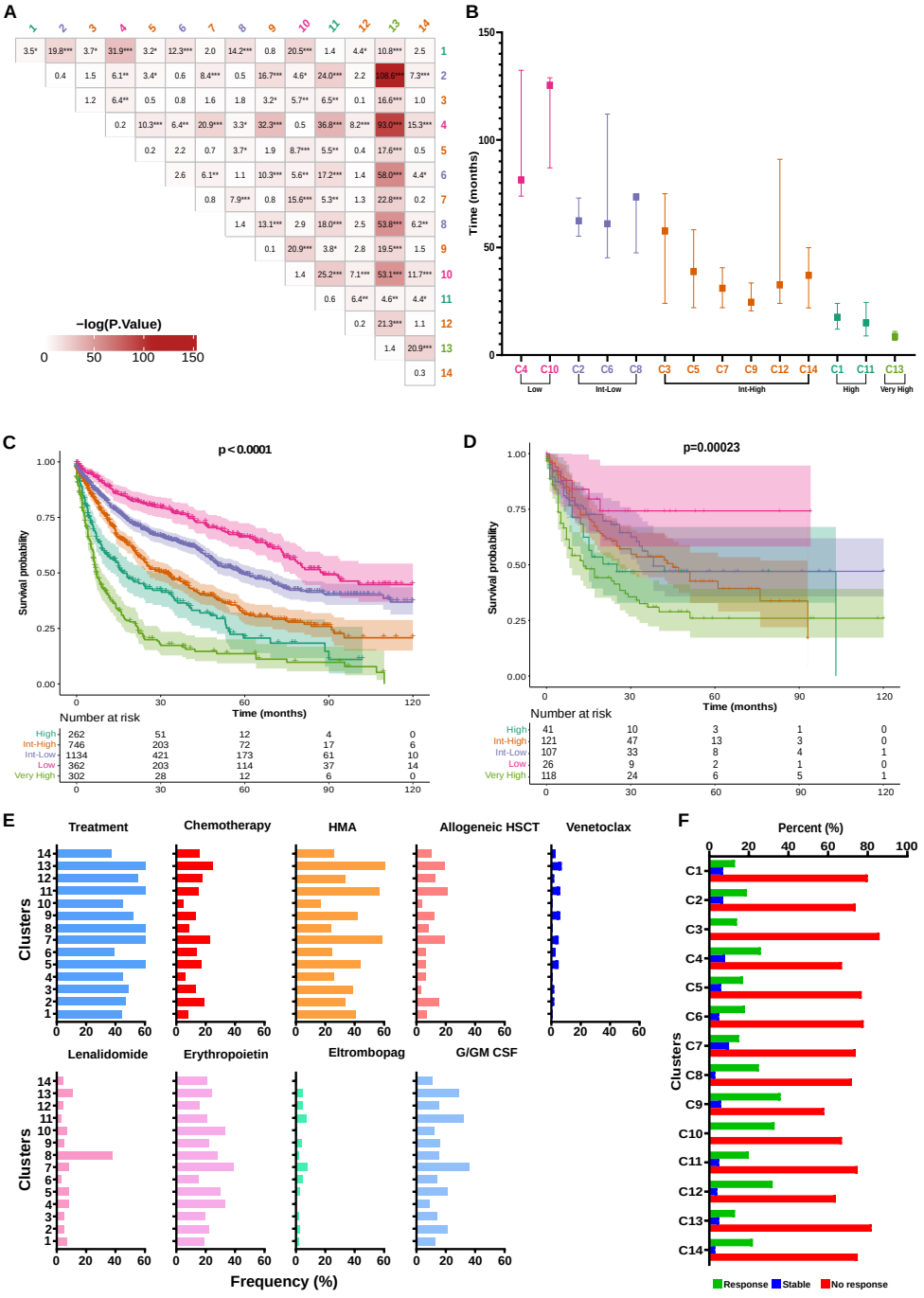


Figure 3.3: **Survival outcomes and model validation. A)** Pairwise survival comparison between the identified genomic clusters. Asterisks indicate the significant -log (P-values). **B)** Median overall survival in months with 95% confidence interval of all molecular clusters and assigned risk groups. **C)** Kaplan-Meier survival curves of all risk groups in the original cohort. **D)** Kaplan-Meier survival curves of all risk groups in the validation cohort. **E)** Bar graph showing the frequency of various first-line treatments used in each cluster. HMA: hypomethylating agents, HSCT: hematopoietic stem cell transplantation, G/MCSF: granulocyte/monocyte colony-stimulating factor. **F)** Histogram bars represent the response to hypomethylating agents treatment among different clusters (C) based on the international working group criteria

Part II

Data Driven Approaches in Solid Tumors

- Published as: **Durmaz A.**, Scott JG. Stability of scRNA-Seq Analysis Workflows is Susceptible to Preprocessing and is Mitigated by Regularized or Supervised Approaches. Evolutionary Bioinformatics. 2022;18. doi:10.1177/11769343221123050

Chapter 4

STABILITY OF SCRNA-SEQ ANALYSIS WORKFLOWS IS SUSCEPTIBLE TO PREPROCESSING AND IS MITIGATED BY REGULARIZED OR SUPERVISED APPROACHES

Statistical methods developed to address various questions in single-cell datasets show increased variability to different parameter regimes. In order to delineate further the robustness of commonly utilized methods for single-cell RNA-Seq, we aimed to comprehensively review scRNA-Seq analysis workflows in the setting of dimension reduction, clustering, and trajectory inference. We utilized datasets with temporal single-cell transcriptomics profiles from public repositories. Combining multiple methods at each level of the workflow, we have performed over 6000 analysis and evaluated the results of clustering and pseudotime estimation using adjusted rand index and rank correlation metrics. We have further integrated neural network methods to assess whether models with increased complexity can show increased bias/variance trade-off. Combinatorial workflows showed that utilizing non-linear dimension reduction techniques such as t-SNE and UMAP are sensitive to initial preprocessing steps hence clustering results on dimension reduced space of single-cell datasets should be utilized carefully. Similarly, pseudotime estimation methods that depend on previous non-linear dimension reduction steps can result in highly variable trajectories. In contrast, methods that avoid non-linearity such as WOT can result in repeatable inferences of temporal gene expression dynamics. Furthermore, imputation methods do not improve clustering or trajectory inference results substantially in terms of repeatability. In contrast, the selection of the normalization method shows an increased effect on downstream analysis where ScTransform reduces variability overall.

4.1 Introduction

Intra-tumor heterogeneity has recently become a central focus of cancer research secondary to the limited response of patients to targeted therapies. These failures are driven by Darwinian evolution, by heritable variation and selection through time. One source for the subsequent intra-tumor heterogeneity is the variation driven by stochasticity in transcriptional activity modulated by epigenetic processes[14, 86]. This change in overall composition is further modulated by the selective advantage of pre-existing resistant cells or clonal expansion of drug-tolerant cells mediated by complex interactions between cells and the microenvironment [87–90]. Although previous efforts have made significant progress in understanding the complex cancer dynamics using bulk sequencing data, single-cell sequencing methods have allowed for novel insights by probing this heterogeneity directly – including during temporally varying processes. For instance, Lee *et al.* identified transcriptional heterogeneity as one of the key factors for promoting the clonal expansion of drug-tolerant sub-population leading to the evolution of resistance [91]. Similarly, Kim *et al.* identified distinct sub-populations resistant to treatment in lung adenocarcinoma patients using single-cell RNA-Seq, [92] Furthermore, relatively recently, single-cell sequencing coupled with mathematical models allowed for investigation of Darwinian dynamics, specifically treatment-induced selection pressure and transcriptional stochasticity at the single-cell level [93–95].

Investigating transcriptional regulation, single-cell sequencing also paved the way for pseudotime/trajectory estimation (PTE) to delineate temporal dynamics during differentiation and resistance evolution. Specifically, PTE aims to find low-dimensional proxy for the underlying transcriptional activity accounting for the temporal information. However, due to the stochasticity inherent in evolution, PTE poses additional challenges where replicate experiments can show divergent dynamics leading to the evolution of distinct resistance mechanisms [33, 96]. For instance, during multipotent progenitor trophoblast differentiation, stages of organization (endpoints) are clearly defined based on morphological charac-

teristics hence we can reliably deduce functional mechanisms through time [97]. However, as we show in detail below, using the same analysis methods with slight differences in pre-processing parameters (number of genes expressed), can result in very different PTE orderings of cells in the setting of the evolution of resistance leading to increased diversity of identified mechanisms. Analysis of single-cell data is further complicated by the technical noise in library preparation strategies due to capture efficiencies at both cell (empty/multi-cell droplets) and transcript level.

In order to alleviate some of the issues with single-cell analysis, various analysis methods aim for robust imputation, outlier detection, and quantification of gene expression. For instance, previous studies utilized imputation methods to reduce the effects of zero-counts due to dropouts in RNA-Seq datasets [25, 26]. In addition to individual methods, multiple packages integrate different analysis stages and tools in unified frameworks; Seurat [98–101], SCANPY [102], BUStools [103, 104]. However, the increased number of available tools, and continued proliferation of them also requires careful selection of methods and associated parameters which can result in significant differences. This issue has been partially addressed before. Specifically, two comprehensive combinatorial evaluation studies have been conducted in order to evaluate different analysis workflows [105, 106]. Tian *et al.* using cell-mixture experiments showed relatively good correlations between ground-truth and estimated trajectories using Slingshot or DDRTree [106]. Similarly, Saelens *et al.* showed improved performance for these methods using topological similarity metrics [105]. While illuminating, a major limitation of these studies is that the methods are applied on non-cancer (embryonic differentiation) processes or cancer cells in relatively homogeneous settings (without selection pressure). For instance, mixture experiments conducted by Tian *et al.* are limited to linear trajectories. In contrast, evolution under selection pressure can result in increased variability and non-linear patterns of transcriptional change [107–109]. As most tumors do not grow in these conditions, it is crucial to evaluate the available methods under selection pressure with temporal information as well. For this purpose,

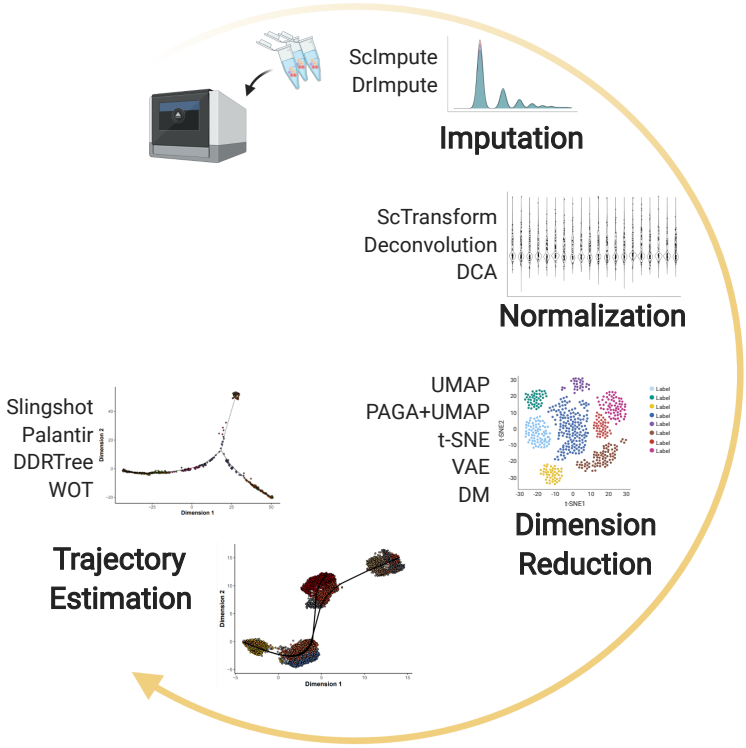


Figure 4.1: **Schematic of general analysis steps and methods used for combinatorial workflows.** Quality filtered raw read counts are processed through a step to reduce possible zero-count inflation by **one of** 2 **imputation methods**; ScImpute, DrImpute (or no imputation). Preprocessed data is **normalized by** 3 **methods**; ScTransform, Deconvolution, and DCA followed by **dimension reduction using** 5 **methods**; UMAP, UMAP+PAGA, t-SNE, VAE, DM. Finally, **one of** 4 **trajectory inference methods** is used; Slingshot, DDRtree, and WOT. Overall we have utilized 6144 analyses for PTE including the data subsets. (Note that the icons representative of individual methods are used to ease the interpretability of combinatorial workflows in downstream figures. Created with BioRender.com)

in this manuscript, we report a benchmarking study in which we evaluate the available methods in a combinatorial fashion similar to Tian *et al.* and Saelens *et al.* focusing on the repeatability of PTEs. We hope that by evaluating the scRNA-Seq methods rigorously for settings applicable to the evolution of resistance in cancer, we will enable more robust and reproducible application of single-cell sequencing technologies and experimental designs for future studies.

4.2 Methods

Single-cell RNA-Seq (scRNA-Seq) analysis follows similar strategies with bulk RNA-Seq where pre-processing is followed by normalization for library size and downstream analysis (see **Fig.** 5.1a for a schema of a typical workflow). Due to the large number of cells being captured non-linear dimension reduction techniques have been extensively used for clustering and trajectory identification such as t-SNE and UMAP [30, 31]. In addition to dimension reduction methods, scRNA-Seq datasets can be zero-inflated due to increased technical noise, hence various imputation approaches have been proposed. Furthermore, a general trend in the scRNA-Seq analysis is to filter out genes that show relatively low variation across the dataset and filter out cells that express a low number of genes. Although this is a valid strategy similar to bulk RNA-Seq analysis, the cutoff for the number of top varying genes to select and the number of genes expressed are generally arbitrary chosen, hence we aim to evaluate the effects of filtering genes and cells based on different thresholds as well. For this purpose, we combine various methods for different levels of analysis in a combinatorial fashion and evaluate identified trajectories in terms of cell orderings (Also note that combinatorial workflows are represented by small icons in downstream figures as column and row labels). Furthermore, since the ground-truth trajectories do not necessarily associate linearly with time in heterogeneous processes (e.g drug resistance), we have profiled clustering performance as well [109]. (See Appendix for a detailed description of methods and parameters)

We have utilized both publicly available datasets and a previously generated in-house dataset with variable number of cells, depth, and complexity of the underlying process (**Table.**4.1). **TKI Treatment** dataset was previously generated to investigate transcriptional dynamics of resistance evolution to 3 Tyrosine kinase inhibitors (TKIs); Alectinib, Lorlatinib, and Crizotinib in lung cancer. To generate scRNA-Seq data with temporal information, cells were sampled at 4h (Alectinib only), 48h, 3w, and 20-24w and sequenced. As we have hypothesized, this dataset represents a biologically heterogeneous example of an evolutionary

process hence crucial to evaluate PTEs. The **Pancreatic cell maturation** dataset contains transcriptional profiles of α and β cells during differentiation process at 7 time-points; embryonic day 17.5 and postnatal days 0, 3, 9, 15, 18, and 60 representing a relatively more homogeneous process with roughly linear sampling times. **Neurodegeneration** dataset is generated to investigate the transcriptional dynamics of microglial cells isolated from Hippocampus at weeks 0, 1, 2, and 6 in CK-p25 inducible mouse model. **E2 Treatment** temporal scRNA-Seq is performed on 2 cell lines (MCF7,T47D) during 17β-estradiol (E2) treatment at 0h, 3h, 6h, and 12h to investigate temporal transcriptional dynamics of estrogen associated pathways in breast cancer. This dataset, however, contains the least number of captured cells sequenced at relatively higher depth.

Each dataset is preprocessed with different gene- and cell-level quality thresholds to generate 12 subsets and the overlap in estimated trajectories are quantified using rank correlation. We have focused on the repeatability of identified PTEs, and aimed to use methods/strategies widely adopted in the community. Additionally we utilize a neural network approach for dimension reduction to evaluate whether more complex models show any advantage when high-throughput single-cell datasets are used. Since neural networks have been extensively utilized for wide variety of problems in the form of autoencoders [110] and relatively recently stochastic alternatives have been used for -omics datasets as well [111–114], neural networks naturally lend themselves to the analysis of single-cell datasets. For comparison of the effect of imputation, we have used ScImpute, DrImpute which showed improved performance in various datasets and Deep Count Autoencoder (DCA) an autoencoder model aiming to combine de-noising and imputation in a single step [25, 26, 115]. We use **2 methods for normalization**; Deconvolution and ScTransform [27, 29]. For DCA, since gene-wise dispersion and mean parameters are already estimated, we only used library size normalization. As scRNA-Seq clustering is an important step utilized in the analysis of various datasets, we wanted to evaluate how robust the clustering results are when different workflows are used as well. For this purpose, we coupled

the Leiden clustering with **5 dimension reduction techniques**; UMAP, PAGA+UMAP, t-SNE, VAE, and Diffusion Maps (DM) and evaluated the overlap of clusters using adjusted rand index (ARI) [30, 31, 116–119]. We have additionally included TooManyCells for clustering, however, due to hardware limitations we used only the Pancreatic Maturation, Neurodegeneration datasets. Furthermore, E2 Treatment dataset resulted in a single cluster across different workflows and subsets possibly due to the low number of cells hence results are not shown [120]. For **trajectory inference,** we evaluated 4 methods commonly used in scRNA-Seq; Slingshot, Palantir, DDRTree and WOT [121–124]. However, Slingshot operates on dimension reduced space hence we combined different dimension reduction methods with Slingshot as well. Palantir in contrast integrates dimension reduction step via diffusion maps to quantify the pseudotime progression from an early cell defined in advance. DDRTree, similarly, generates cell orderings by reducing the high-dimensional data to low-dimensional principle-tree structure, hence we have coupled DDRTree with preprocessing and normalization steps only. Furthermore, since Slingshot and DDRTree are unsupervised approaches, we have utilized Waddington-OT (WOT), a supervised approach that aims to find cell-cell transition probabilities at consecutive time-points via optimization of unbalanced transcriptional mass transfer. Comparison is somewhat imperfect however, as trajectories are defined slightly differently for each method. Since Slingshot estimates the smooth principle curve in low dimensional space, mapping of individual cells on the curve readily defines an ordering via the arc-length along the curve. In contrast, DDRTree embeds high-dimensional transcriptomic profiles onto a principle tree structure where the ordering is defined by the geodesic distances between individual cells. The supervised approach, WOT, on the other hand, generates a probability distribution between an individual cell at time t_i and the cell population at time t_{i+1}, hence the trajectory of an individual cell is defined as the vector of transition probabilities. In order to evaluate the results from different methods in a comparable fashion, we opted to use Spearman's ρ which does not take into account the distances between individual cells, but rather only the orderings, hence

Dataset	Subset	Size (#cells/#genes)	Platform
TKI Treatment [125]	Alectinib	5000/14000	10x
	Lorlatinib	4000/14000	10x
	Crizotinib	3700/14000	10x
Pancreatic Maturation [126]	α cells	250/21700	SmartSeq2
	β cells	410/20500	SmartSeq2
E2 Treatment [127]	MCF7	60/21395	Fluidigm C1
	T47D	60/21570	Fluidigm C1
Neurodegeneration [128]	–	800/15545	SmartSeq2

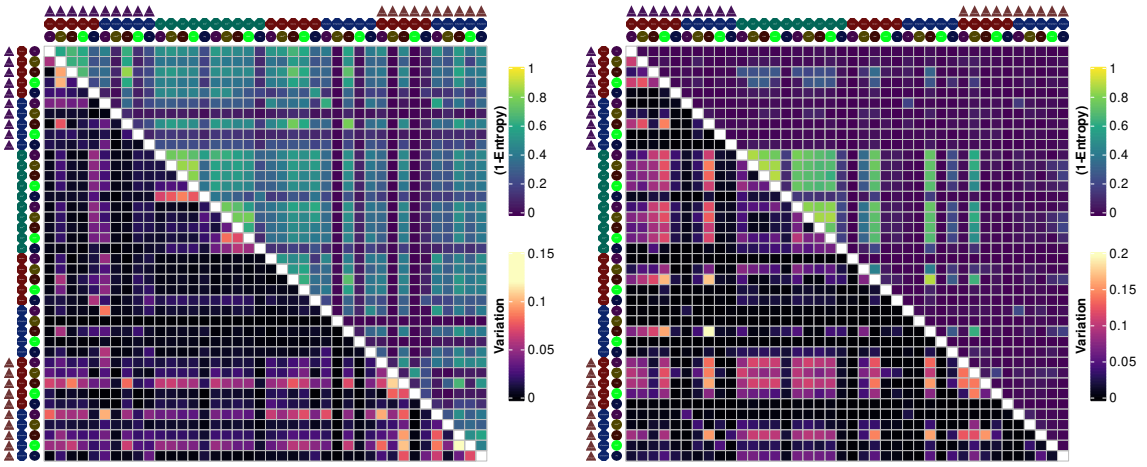
Table 4.1: Datasets utilized in the study where the number of cells and genes are given prior to subset generation after quality control

different quantitative scales between methods can be compared.

4.3 Results

4.3.1 Dimension reduction & Clustering

In order to evaluate how dimension reduction methods perform when coupled with the Leiden method for clustering we have compared the identified clusters using Adjusted Rand Index (ARI) across different subsets of gene and cell level thresholded datasets. However, note that since we do not have ground-truth observations of clusters, instead we have focused on the overlap of identified clusters between different methods to assess repeatability. Specifically, individual dimension reduction methods coupled with different preprocessing steps (imputation and normalization) are used to generate clustering via the Leiden method. Generated individual clusters are then compared using ARI and ARI values across different subsets are aggregated by taking the median of ARI values. This approach allowed us to investigate the stability of clusters for a given dimension reduction method when combined with different pre-processing steps. Furthermore, a common practice in scRNA-Seq analysis is preprocessing with Principal Component Analysis (PCA) to both reduce computational load and reduce variation/noise which requires selection of number of *top* latent features to keep where automated tools can be utilized [129]. However, dimension reduction via PCA can be non-trivial and introduce unwanted bias specifically in the case of multiple datasets hence we opted to not utilize PCA as an initial preprocessing step. As expected we observed



(a) Pancreatic differentiation α cells (b) Pancreatic differentiation β cells

Figure 4.2: **Comparison of trajectories identified by Slingshot showing data dependent performance of each workflow.** Combinations of icons for columns/rows represent distinct workflows. Entropy (upper triangle) is used to aggregate over multiple trajectories identified by Slingshot and data subsets corresponding to cell level and gene level filtering thresholds. Variation (lower-triangle) over different data subsets is given to show the confidence for aggregating Entropy measure (See Supplementary for details). Results suggest data dependence where the use of imputation in β cells dataset significantly reduces the overlap of PTEs in contrast imputation step overall preserves the identified PTEs in α cells.

a positive correlation of ARI across different workflows with the number of cells (**Fig**. C1). However, ARI values showed reduced overlap between different methods across datasets globally, even when the number of cells is high ($ARI < 0.75$). Investigating methods individually showed t-SNE as relatively more robust to different preprocessing steps in the TKI dataset where remaining datasets showed variable performance (**Fig**. C2a,C2b,C2c). Interestingly, neural-network methods showed variable results where the use of DCA-NB/DCA-ZINB as a preprocessing step in the TKI dataset led to improved overlap between UMAP, UMAP+PAGA, and t-SNE. In contrast, the use of VAE as a dimension reduction method showed poor performance resulting in variable cluster assignments (See **Fig** C3 for example workflows). This suggests that as a dimension reduction method, neural-networks might not be the optimal choice but as a preprocessing step neural networks can provide advantages depending on the number of cells.

Datasets with relatively low number of cells showed variable results. For instance, in

Pancreatic α cell differentiation, UMAP and PAGA+UMAP showed improved overlap when DrImpute is combined with Deconvolution but the overlap is reduced when ScTransform is used for normalization (**Fig.** C2f). The use of t-SNE similar to the TKI dataset was more robust to preprocessing steps. E2 Treatment dataset resulted in variable cluster assignments overall where both MCF7 and T47D cell line datasets resulted in different cluster assignments across workflows. The Neurodegeneration dataset on the other hand benefited from DCA with or without zero-inflation model but overall showed decreased overlap as well (**Fig.** C2h).

To further extend the analysis results, we have evaluated tooManyCells method as well which is another scRNA-Seq method used for clustering nearest-neighbor graphs to partition the cell population [120]. tooManyCells improved cluster overlap globally in the Pancreatic Maturation and Neurodegeneration datasets (**Fig.** C4). However, similar to Leiden clustering, selection of preprocessing workflows showed data-specific performance. For instance, the Pancreatic maturation α cells dataset was more sensitive to the imputation with DrImpute in contrasts with β cells dataset where imputation with DrImpute showed reduced overlap in cluster assignments when ScTransform is used for normalization (**Fig.** C4a). Interestingly, in the Neurodegeneration dataset, a dichotomy between the use of ScTransform and other workflows is observed where ScTransform showed poor overlap with other workflows (**Fig.** C4b).

We have also investigated the overlap of identified clusters with temporal information. Specifically, using homogeneity metric via R package *clevr*, we quantified the distribution of cells sampled from different time-points in a given cluster in order to delineate whether given methods can distinguish cells from different time-points. We observed a general improvement when TKI dataset is considered specifically when t-SNE, UMAP or PAGA-UMAP is applied where interestingly E2 treatment dataset showed the lowest homogeneity(**Fig.** C19). Furthermore, when ScTransform is coupled with DrImpute, substantial decrease in Pancreatic Maturation and Neurodegeneration datasets is observed which sug-

gest that workflow selection should be done in a data specific fashion.

4.3.2 Trajectory Estimation

In order to evaluate PTEs mapping to a latent biological process we used Spearman rank correlation and normalized entropy. As given previously, using rank correlation we aim to do a comparison of cell orderings identified by different workflows and normalized entropy is used to assess the distribution of rank correlations (bimodal around 0-1) in the case of Slingshot since > 1 PTEs are identified (**Fig.** C5).

4.3.2.1 Slingshot

Evaluating the trajectories identified by Slingshot, we have observed large variation across different workflows and across different subsets. For instance, in Pancreatic maturation datasets, workflows that show relatively good overlap in α cell trajectories failed to identify overlapping trajectories in the β cell dataset. Specifically, the use of DrImpute or ScImpute resulted in decreased overlap of PTEs in β cell dataset (**Fig**. 4.2). Furthermore, the number of cells did not correlate positively with the repeatability of identified trajectories where the majority of the workflows showed high entropy of rank correlations in the TKI treatment dataset with minimum entropy being > 0.7 across 3 treatments (**Fig**. C6a,C6b,C6c). In contrast, datasets with relatively low number of cells showed slightly improved overlap for specific workflows. For instance in the E2 treatment dataset, use of DM improved overlap in contrast with UMAP or UMAP+PAGA. The Neurodegeneration dataset on the other hand showed a global decrease in PTEs (**Fig**. C6).

These results point out one of the major drawbacks of using Slingshot for PTEs; since the estimation of trajectories is heavily dependent on the prior dimension reduction step, heterogeneous datasets will necessarily show high variation to different parameter regimes. More specifically, the Slingshot method using principle curves can fail to capture the temporal dynamics on highly non-linear spaces hence need to be carefully selected/optimized

for trajectory estimation. For instance when UMAP is used for dimension reduction prior to PTE, the cell population structures remain overall similar as the number of cells increases but relative positioning of subpopulations can change in a way that does not reflect the latent temporal process (**Fig**. C7a,C7b,C7c,C7d). Furthermore the non-linearity can artificially generate an increased number of trajectories resulting in diverge PTEs. For instance, use of DM resulted in 1 trajectory to be identified in E2 treatment data subsets hence resulting in 'simpler' PTEs overall (**Fig**. C7g,C7h).

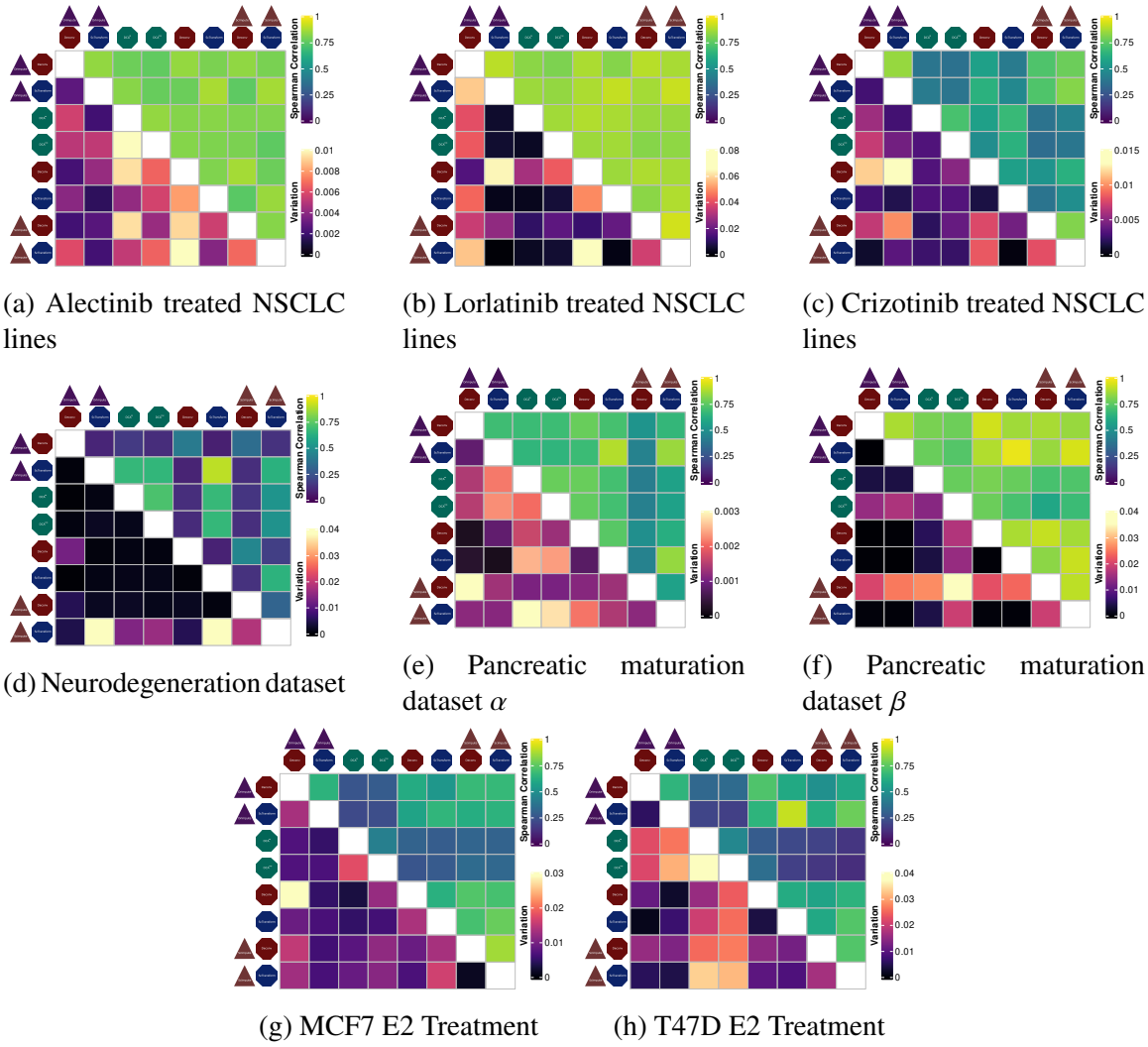
4.3.2.2 Palantir

We have also included an additional method widely used for pseudotime estimation[122]. Palantir utilizes nearest-neighbor graphs followed by diffusion maps as a dimension reduction/manifold learning step. Low dimensional representation is further used for pseudotime quantification as a distance measure from a defined progenitor cell. In order to marginalize out the selection of progenitor cell, we generate 10 pseudotime orderings using different progenitor cells sampled from initial time-point and calculate the average Spearman rank correlations. In the TKI dataset, Palantir showed relatively robust estimates of pseudotime orderings across different preprocessing steps where the average correlation remained > 0.5 (**Fig.** C8). However, similar to Slingshot results, data-specific overlap quality was present. For instance, the Alectinib treated dataset benefited from imputation by ScImpute across different subsets but Crizotinib and Lorlatinib treated datasets showed reduced overlap of PTEs. Furthermore, Crizotinib and Lorlatinib treated datasets showed distinct profiles for DCA-NB/DCA-ZINB where Crizotinib dataset benefited across different subsets from using DCA but Lorlatinib dataset showed subset dependent profile. In the Neurodegeneration and Pancreatic Maturation datasets, similar results were observed where ScTransform normalization helped improve the PTE overlap in the Neurodegeneration dataset but showed reduced overlap in the Pancreatic Maturation dataset specifically when coupled with DrImpute. Conversly, using DCA-NB/DCA-ZINB, Palantir PTEs showed relatively robust

correlation across different workflows in the Pancreatic Maturation dataset. (**Fig.** C9,C10).

4.3.2.3 DDRTree

Since DDRTree/Monocle2 method inherently utilizes dimension reduction to generate a tree-like topology to define a latent trajectory, we have generated the combinatorial workflows for imputation and normalization steps only which is also reflected on the use of 2 icons instead of 3 where imputation is applied. However, also note that, in contrast with previous workflows, we have opted to further reduce the number of features by selecting top 50 principal components due to computational constraints hence the limitation of results to within method comparisons. We have aggregated rank correlations across 12 subsets by median values to evaluate the overlap of different workflows. (**Fig**. C11). The TKI treatment dataset overall showed good overlap ($\rho > 0.75$) across different imputation and normalization methods. Interestingly however, Crizotinib treatment showed increased overlap of PTEs when DrImpute or ScImpute is utilized in comparison with when DCA is used (**Fig**. 4.3a,4.3b,4.3c). Further investigating the individual trajectories showed that using DCA resulted in increased number of branch points in contrast with DrImpute or ScImpute (**Fig**. C12). This might be an implication for 'overcorrection' when DrImpute or ScImpute is used subsequently reducing variation. Datasets with relatively low numbers of cells however showed variable results with different analysis steps having distinct 'importance'. For instance, in the Neurodegeneration dataset, choice of normalization showed the highest impact where the use of Deconvolution decreased the trajectory overlap globally, in contrast, ScTransform was more robust to the imputation step (**Fig**. 4.3d). Furthermore as expected, E2 treatment dataset showed high correlation between workflows using Deconvolution and ScTransform normalization but not when DCA is used (**Fig**. 4.3h, 4.3g) since, with low number of training dataset for the autoencoder model, parameters might not be estimated robustly. In contrast, pooling information from similar cells and genes might better capture biological signal. The Pancreatic differentiation dataset on the other hand showed increased



(a) Alectinib treated NSCLC lines

(b) Lorlatinib treated NSCLC lines

(c) Crizotinib treated NSCLC lines

(d) Neurodegeneration dataset

(e) Pancreatic maturation dataset α

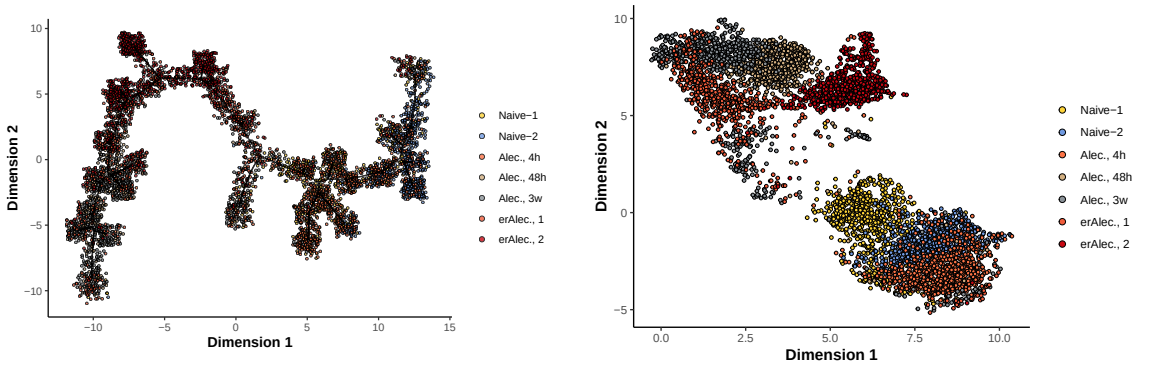
(f) Pancreatic maturation dataset β

(g) MCF7 E2 Treatment

(h) T47D E2 Treatment

Figure 4.3: **Rank correlation of geodesic distances on DDRTree trajectories median aggregated over subsets showing both data specific performance and overall increase based on number of cells.** (a-c) TKI treatment dataset shows improved overlap of cell orderings. Altough the TKI dataset is relatively more heterogeneous, increased number of cells allow DDRTree to capture robust cell-cell similarities. (d-h) Remaining datasets show variable results with Pancreatic maturation β performing comparable to TKI dataset and Neurodegeneration dataset performing the poorest.

overlap across different methods. Further investigating subset specific overlap of trajectories showed no substantial effect of gene or cell level quality filtering where the quality of overlap between different workflows remained similar across different subsets (**Fig**. C13).



(a) DDRTree

(b) PAGA-UMAP

Figure 4.4: **Sample dimension reductions for Alectinib treated NSCLC lines showing nonlinearity in temporal dynamics of gene expression.** Since dimension reduction utilizes transcriptional similarity of individual cells, low dimensional representations might not necessarily correlate linearly with sampling time. In datasets where sampling time is not linear and/or the underlying transcriptional dynamics are highly heterogeneous supervised approaches might be more suitable where the change in transcriptional activity is ordered by the temporal process by default.

4.3.2.4 WOT

Since both Slingshot and DDRTree aim to find a low dimensional ordering of individual cells in an unsupervised fashion, temporal information is not readily utilized which can lead to biased estimates where transcriptional dynamics are not 'linearly' associated with time (**Fig.** 4.4). Instead, supervised approaches can provide certain advantages for PTE by utilizing available temporal information. However, forcing individual cells in a supervised order also poses challenges such that cells are not synchronized in terms of division and growth rates. For this purpose the WOT framework also allows us to calculate optimal growth rates for individual cells given the 'transcriptional mass' transfer optimization problem. Furthermore by removing the dimension reduction step, WOT inherently reduces the number of possible sources of variation. In order to evaluate how WOT performs when different methods for imputation and normalization are used, we have calculated pairwise rank correlations of transition probabilities between individual cells at consecutive time-points t_0, t_1, across different workflows. Simply, we have quantified how the transition probabilities of an individual cell change if a different normalization or imputation step is used. The TKI

treatment dataset showed the highest overlap of transition probabilities across all pairwise workflow comparisons with median rank correlation > 0.75 (**Fig**. C14). Normalization with ScTransform showed slightly better overlap however when different imputation steps are compared in contrast with Deconvolution (**Fig**. C15). Interestingly however this difference was most striking in the Neurodegeneration dataset where the choice of imputation showed a relatively high difference of rank correlation ($\rho > 0.2$) between Deconvolution and ScTransform. Investigating imputation steps individually showed no substantial effect of imputation step where the overlap of trajectories when Deconvolution and ScTransform is used remained similar and relatively low (< 0.75) irrespective of which imputation step is used (**Fig.** C16). Investigating the effect of using different gene and cell level thresholds showed a substantial decrease in the Neurodegeneration dataset where the remaining datasets showed similar PTE comparison profiles across 12 subsets hence suggesting relatively robust PTEs across different threshold (**Fig.** C17). This suggests that WOT PTEs consistently show repeatable results specifically for datasets with relatively high number of cells captured (**Fig.** C18).

4.4 Discussion

With the advent of single-cell sequencing methods, identification of tumor subpopulations and pseudotime estimation has been extensively used where analysis of scRNA-Seq data is complicated by a multitude of factors. In order to evaluate methods developed for scRNA-Seq analysis we have aimed at evaluating the available methods in a combinatorial fashion to assess the repeatability of either identified subpopulations or estimated pseudotimes. We have shown that selection of different methods at different levels of scRNA-Seq analysis can lead to variable outcomes both for clustering and trajectory inference. This is especially important considering the availability of additional methods not utilized in this study and the continued proliferation of methods [130, 131]. Furthermore, we have observed substantial variation in workflows for either clustering or PTE where non-linear dimension reduction

methods are used. This emphasizes the importance of careful evaluation of which methods to utilize since the results may not be generalizable to replicate datasets.

General trends in our analysis showed that the number of captured cells is crucial when deciding on which downstream analysis methods to use since datasets with relatively high number of cells can sample the evolutionary process on the underlying manifold more effectively hence showed increased overlap across different workflows specifically for clustering and PTE using WOT. Imputation approaches did not show improvement in downstream analysis as well which have been previously reported as well [132]. Dimension reduction methods that are heavily utilized in scRNA-Seq analysis showed high sensitivity to parameter selection hence clustering results using low dimensional representations were variable. Similar results were also shown previously [133]. Although, t-SNE and UMAP coupled with PAGA showed relatively robust cluster assignments there is no one best approach and methods showed data-specific performance. Clustering with tooManyCells on the other hand alleviated some of the limitations where clustering is done via nearest-neighbor graphs, however, data-specificity of workflows remained. This further stresses the importance of repeatability in scRNA-Seq analysis where unsupervised clustering is of major interest. In order to reduce some of the issues associated with clustering specifically when coupled with non-linear dimension reduction, 'consensus' based approaches where randomly sampling features/cells might be more suitable.

Trajectory inference methods, similarly showed variable results where non-linear dimension reduction is used. Slingshot for instance failed to capture reproducible trajectories in the TKI treatment dataset. As previously stated, Slingshot method based on principle curves is more suitable to relatively linear trajectories with a small number of branch points. Similar observations were also pointed out in previous studies as well. For instance, *Saelens. et al.* showed decreased performance of Slingshot when the underlying trajectory consisted of multiple branches and non-linearity, which as we have shown in this study, can be further exacerbated when considering multitudes of different preprocessing steps. In order to

alleviate some of the issues in coupling dimension reduction methods with Slingshot, one may need to choose parameter regimes towards linearity, for instance increasing number of nearest neighbors or minimum distance parameter in UMAP. However, also note that the use of dataset from a single experimental setup is of limited applicability hence does not necessarily dismiss alternative views in the case of the TKI treatment dataset. Palantir on the other hand resulted in more robust PTEs across data subsets. This can be attributable to the fact that Palantir readily optimizes the number of dimensions to use to quantify PTEs hence reducing variation overall. Nevertheless, Palantir also suffered from data-specificity. Using either supervised approaches WOT or regularized dimension reduction using DDRTree resulted in increased correlations in trajectory estimates when different preprocessing methods are combined. DDRTree specifically showed improved performance over Slingshot especially when ScTransform is used for normalization but the quality of the overlap was data specific where TKI dataset with relatively large number of cells showed a global increase in correlation of identified trajectories. This is in contrast with *Tian et al.* where Slinghost showed slightly improved performance over DDRTree. However, improved performance of Slingshot can be partially attributed to the mixture datasets being relatively less heterogeneous and the underlying structure being relatively linear.

Using supervised trajectory mapping via the WOT framework alleviated some of the issues with unsupervised approaches as well. Although identified trajectories remained sensitive to normalization method selection, data dependence is reduced where we have observed ScTransform performing relatively well across all the datasets. Furthermore, since the temporal information is utilized in WOT, we can readily assume the identified trajectories will overlap with the biological process compared to unsupervised alternatives. For instance, neither Slingshot nor DDRTree can differentiate subpopulations from different time-points if the transcriptional profiles are similar even though the temporal dynamics are different. However, it is also important to note that identified trajectories only regard the differences between individual cells in terms of transcriptional profiles mapped to low dimensional

space (in the case of Slingshot and DDRTree). This makes the problem of evaluating the PTEs non-trivial due to absence of ground-truth observations Deviation from ground-truth PTEs should be evaluated using approaches that allow individual cells to be tracked [134, 135]. Furthermore, individual methods presented here can be further optimized separately resulting in improved PTEs. For instance, increasing the number of dimensions or using alternative metrics for quantifying transcriptional difference. Nevertheless, the WOT framework combined with ScTransform provided certain advantages by utilizing temporal information and reducing the variation.

In conclusion, analysis of scRNA-Seq datasets show high variation across different parameter regimes and methods in the context of clustering and trajectory mapping. It is non-trivial to utilize the heterogeneous structure of tumor subpopulations in order to extract biological insights hence analysis of scRNA-Seq requires careful selection of methods and optimization of parameters but different methods provide certain advantages. We hope that provided results can guide future studies for method selection and help with reproducibility in scRNA-Seq analysis.

Chapter 5

PANCANCER MAPPING OF COLLATERAL SENSITIVITY USING MULTI-OMICS ML APPROACH

Evolution of drug resistance is a major obstacle in cancer treatment where both targeted and non-targeted agents often fail to deliver complete cure. Recent efforts to understand the underlying dynamics based on mathematical modeling have allowed for better treatment strategies prolonging progression-free and/or overall survival. Concurrently, collateral-sensitivity has been defined and elaborated. However, improved mechanistic understanding of individual drug effects to better understand collateral sensitivity/resistance remains to be explored.

Here we investigate the utility of integrative machine-learning approach based on autoencoders in capturing functional mechanisms of drug sensitivity/resistance. We pose this problem as finding a low-dimensional landscape/embedding capturing covarying features across gene-expression and mutation profiles supervised by drug sensitivities. Furthermore, we train the proposed model simultaneously on patient and cell-line datasets to extract clinically relevant features. We use bulk RNA-Seq and whole-exome mutation data from TCGA and GDSC in a pancancer fashion. We further utilize IC50 measurements for 120 drugs to supervise the landscape. Filtering the features to focus on protein coding and driver genes for expression and mutation profiles respectively, we generate an integrative map of 10k gene expressions, 500 driver genes associated with 120 drugs across patient and cell-line datasets. We show the capability of such models in integrating multiple -omics datasets to uncover potential convergent mechanisms of resistance/sensitivity in a pancancer fashion and evaluate both on survival predictions setting and single-cell RNA-Seq resistance evolution setting.

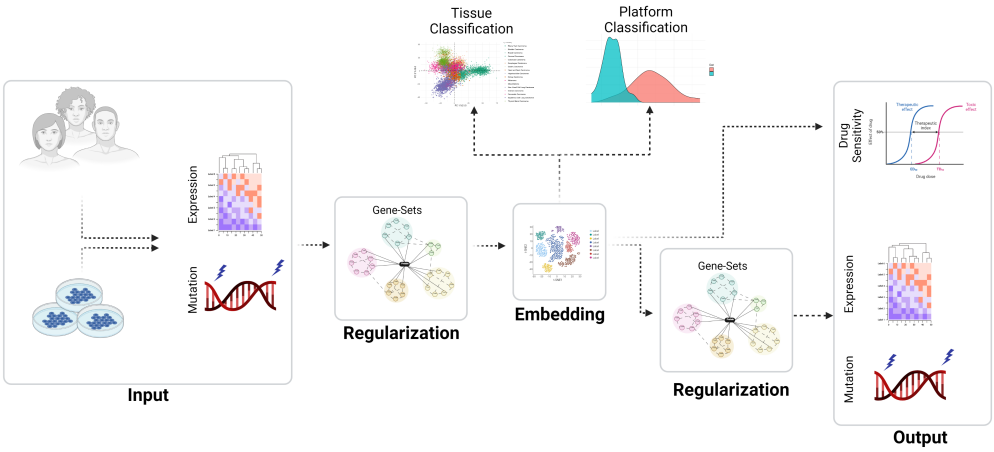
5.1 Introduction

Resistance to targeted and cytotoxic agents and subsequent proliferation is inherent in cancer where evolutionary dynamics are at play. Through selection of preexisting clones, *de-novo* emergence of resistance and/or ecological interactions with the tumor microenvironment, cancer cells can adapt/resist to environmental perturbations making it non-trivial for complete remission specifically for advanced stage/metastasized cases. Owing to the increasingly available -omics datasets including single-cell sequencing and development of mathematical frameworks, better understanding of the evolutionary dynamics have resulted in improved treatment strategies [35, 37]. Furthermore, with the aim of preventing resistance evolution, large-scale drug combination studies have been conducted to delineate synergism/antagonism of anti-cancer drugs for increased effectiveness and reduced toxicity [136, 137]. Coupled with off the shelf black-box models, drug combination studies allowed for effective prediction of synergism/antagonism admittedly with poor translation to clinic [138–140].

Relatively recently, evolutionary view of drug sensitivity through mathematical modeling, linked collateral mechanisms of sensitivity/resistance with fitness landscapes in bacteria and in cancer further elaborating on the stochastic nature of evolutionary mechanisms of drug resistance [33, 36, 38]. A significant difference of collateral sensitivity over static drug-combination view is that the identification of temporal information via collateral networks can in theory allow for better control of drug resistance through time, in contrast the intrinsic adaptability/plasticity of cancer cells to environment allows for evolution of resistance to drug-combinations given in pairs. Collateral-networks however require mechanistic view of individual drugs in order to be able to administer drugs temporally with non-orthogonal mechanisms of resistance.

For this purpose, here, we aimed at mapping latent collateral sensitivities of anticancer drugs via the use of neural-network based machine-learning models. Specifically, we pose the problem as a supervised low-dimensional embedding and subsequent prediction of drug

sensitivities simultaneously to multiple drugs. We generate an integrative model based on autoencoder architecture to identify latent features associating multi-omics features with drug sensitivities. Furthermore, we leverage patient data from TCGA as well, regularizing the identified features to be generalizable to the clinical setting.



(a)

Figure 5.1: **Integrative approach capturing covarying features of mutation and gene expression associated with drug sensitivities through a bottleneck:** Autoencoder architecture representing the supervised integrative embedding approach.

5.2 Results

5.2.1 Integrative approach can provide mechanistic view of collateral sensitivity

We have constructed an Autoencoder architecture in order to identify covarying features across different -omics types and datasets where data is embedded on a unified space effectively compressing relevant information (**Fig.5.1a**). The proposed model is further supervised by drug sensitivities measured by IC50 to enable capturing information associating expression and mutation patterns across multiple drugs (See methods for more details).

Using cross-validation to select the number of latent dimensions, we observed relatively high overlap of predictions for test dataset both for patient and cell-line samples where mutation predictions on average showed > 0.7 and > 0.9 AUC and expression predictions on average showed > 0.65 and > 0.80 pearson correlation for cell-line and patient datasets

respectively (**Fig.D2**). However, drug sensitivity predictions, although being relatively high in the training set ($\rho > 0.75$), showed reduced overlap in the test set ($\rho > 0.4$) (**Fig. D3**). Nevertheless, we have observed pathway and cancer specific improvement in predictions. For instance, drugs targeting ERK/MAPK signaling, Mitosis and DNA replication showed $\rho > 0.5$. Similarly, Ewing's Sarcoma, Hepatocellular Carcinoma and Small Cell Lung Carcinoma showed $\rho > 0.5$ in test-datasets. Cancer and pathway specificity was also apparent when investigating the rank correlations in the training dataset where majority of the cell-lines showed $\rho < 0.4$ absolute correlation when considering all the drugs in the dataset (**Fig. D4**). This observation underlines the sparse nature of drug response mechanisms which can be alleviated by 'borrowing' power from drug-cancer type combinations with relatively high mutual information.

5.2.2 Low dimensional latent space recapitulates feature associations

In order to investigate the linear associations between the features and drug sensitivities, we generated samples from the latent space and quantified the linear associations. Identified associations recapitulated previously known mechanisms or signatures (*e.g.* *BRAF* mutation conferring sensitivity to Refametinib, *KRAS* mutation conferring resistance to Gefitinib and Cisplatin sensitivity signature showing positive association with previously published gene expression signatures [141] (**Fig.D5**)). Interestingly, gene-set enrichment analysis showed a global association of chromatin organization, and histone-deacetylation through *ZZZ3* (**Fig.D9**) where histone deacetylase inhibitors (HDACi) have been previously implicated for increased synergistic effects in a class dependent fashion as well and is suggestive of a possible convergent pathway in drug response [142–144].

However, overall, we have observed high frequency of positive correlations (reduced collateral sensitivity) in drug-drug similarities (**Fig. D6**). Comparable results have been given previously as well in the context of synergistic drug combinations and collateral sensitivity emphasizing the sparsity of such combinations [137]. In contrast, a handful of drugs

targeting *ERK/MEK, EGFR*, heat-shock proteins and intrinsic apoptosis pathway including Refametinib, Trametininb, Sapitinib, Sepantronium bromide, Navitoclax and Tanespimycin (**Fig. 5.2**).

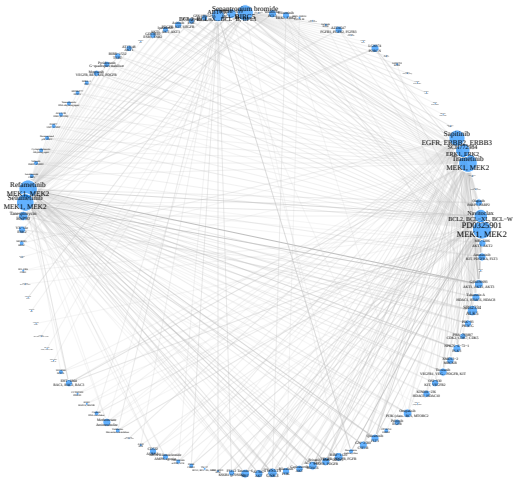


Figure 5.2: **Drug-Drug network showing drugs with negative association and $\rho > 0.4$ test-set prediction correlations**.

In order to investigate the possible mechanisms of negative associations, we used biased random-walks coupled with high-confidence gene-gene interaction network aggregated from StringDB, BioGrid and InBioMap. Using features with high 'loadings' on top 5 singular vectors of feature-drug correlation matrix as restart nodes/seeds, we identified sub-networks of possible synergistic/collaterally sensitive drug mechanisms (See supplementary for details). Identified subnetworks for top 5 components included; *BRAF,BCOR,CDKN2A,S100A10,S100A VIM,NOTCH1,NRAS,TET1,SYK* component, *TP53,DUSP6,PAX6,NGFR,BRAF* component, *HLA-B, CPEB2* component and *PTEN,CDKN2A,GAB1* (**Fig. D7**).

Further investigating the feature-drug associations for *all* the drugs with test-set predictions $\rho > 0.4$, we observed possible collateral sensitivity/synergy between Navitoclax a *BCL2* inhibitor and Gefitinib a *EGFR* inhibitor in a cancer specific fashion. Small Cell Lung Cancer, Neuroblastoma, Ewing's Sarcoma cell lines were sensitive to Navitoclax and resistant to Gefitinib whereas Cervical Carcinoma, Esophageal Squamous Cell Carcinoma, Head

and Neck Carcinoma and Oral Cavity Carcinoma were resistant to Navitoclax but sensitive to Gefitinib (**Fig. D8**). Interestingly a *HSP90AA1* inhibitor Tanespimycin which has been previously implicated as a potential target for TKI resistant Non-Small Cell Lung Cancer NSCLC [145, 146], showed (-) correlation with Navitoclax as well. In line with previous publication showing *HSP90AA1* inhibition promoting apoptosis by blocking pro-survival signals through *AKT* [147], we hypothesized that Gefitinib and Tanespimycin resistance, is dependent on upregulation of pro-survival signals leading to Navitoclax sensitivity.

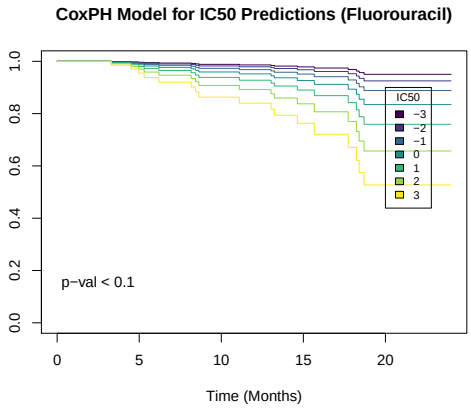
Consequently, stimulating further the intrinsic apoptotic pathway by inhibition of additional pro-survival mechanisms can improve both the efficiency of treatment and possibly restrict the diversity of resistance mechanisms that can evolve. Improved efficiency has been previously shown in NSCLC treated in combination with *EGFR* inhibitors Erlotinib, Gefitinib and *BCL2* inhibitors [148, 149].

In order to investigate the dynamics of *EGFR,HSP90AA1,BCL2* inhibition, we quantified the sensitivity of Lung Cancer cell-line (PC-9) harboring deletion of exon 19 in *EGFR* to Navitoclax and Tanespimycin. We tested the combination in both parental PC-9 cells and Gefitinib resistant PC-9 cells evolved over 6 months with increasing dose of Gefitinib. Interestingly we observed synergy in both parental and resistant cell-lines, however, parental cell-line showed a dose dependent synergy with increasing Navitoclax concentration. In contrast, Gefitinib resistant cell-lines showed synergy across all concentrations of Navitoclax suggestive of *BCL2* dependent resistance evolution to Gefitinib (**Fig. D10**).

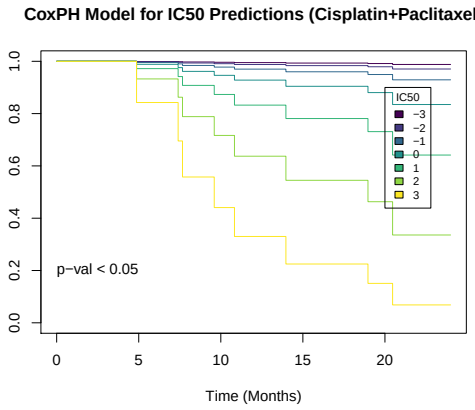
5.2.3 Drug sensitivity predictions show significant association with progression-free survival

Since drug sensitivity measurements are aimed at identifying tumor populations which a specific drug will be most effective, IC50 can also be a proxy for how well a given patient will respond to treatment. With the naive assumption of patients with low predicted IC50 values should respond better to that treatment, we investigated the utility of the trained

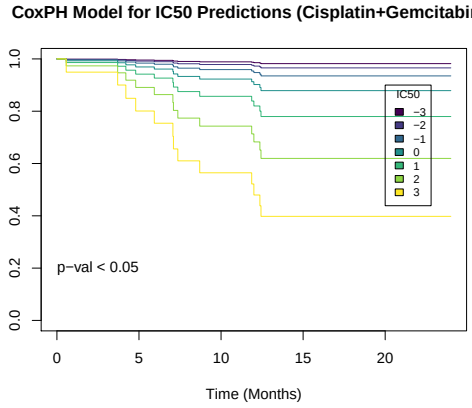
model in predicting progression free survival. We curated TCGA clinical information and extracted stage, radiation and drug treatment information and combined with a published time-to-event dataset [150].



(a)



(b)



(c)

Figure 5.3: **CoxPH survival analysis using progression free survival and predicted IC50 values for the corresponding drugs or mean aggregated values for multiple drugs**. Shown survival curves represent the effect of IC50 predictions of a pseudo-samples of age 45, low stage (where applicable) and radiation treated (where applicable).

Using the predicted IC50 values as covariates in CoxPH model, we identified 3 cancer types (out of 6 applicable cancer type-drug combinations) and treatments with significant association with IC50 predictions for BLCA, OV and STAD (**Fig. D13**). However, remaining combinations including cancers CESC, HNSC, LIHC did not show significant association. This might be due to IC50 predictions not-necessarily translating into survival phenotype

or the heterogeneity of clinical data curation regarding surgery, dose of radiation etc.. Nevertheless, in a limited setting, we showed clinical utility of the trained model.

5.2.4 Drug specific application of the proposed Autoencoder model

Hypothesizing that pancancer analysis of the observed variation can mask tissue specific sensitivity patterns of given drugs we investigated whether limiting the neural-network to train on specific drugs could improve predictive power of the model hence we restricted our analysis to 4 drugs including targeted agents Alectinib, Crizotinib, Osimertinib, Erlotinib and evaluated on single-cell RNA-Seq data obtained from public repositories [125, 151].

scRNA-Seq datasets are generated to measure transcriptional activity during resistance evolution. For that purpose, dataset (1) uses EML4-ALK+ isogenic cell lines treated with Alectinib, Crizotinib and Lorlatinib over 6 months to develop resistance at clinically applicable doses. Dataset (2) instead uses *EGFR*+ cell lines treated with Erlotinib over 2 weeks and sampled at relatively shorter intervals. Similarly dataset (3) uses PC-9 cell lines treated with Osimertinib and sequenced at 2 time-points; parental and day 5 with multiple evolutionary and technical replicates. However, since we only had gene expression measurements, we retrained our model with expression dataset only and restricting the genes to a common set of 6k genes expressed in TCGA, GDSC bulk RNA-Seq and scRNA-Seq datasets.

Since the scRNA-Seq samples are proxy for resistance development, we reasoned that increase in predicted IC50 values should correspond to later time-points where cells have developed resistance. Indeed we observed a significant increase in predicted IC50 values at later time-points for *EML4-ALK*+ dataset (**Fig. D11a**). Furthermore, the predictions overlapped with observations of cells treated with Lorlatinib as well, where cells treated with Lorlatinib showed higher resistance to Crizotinib compared to cells treated with Crizotinib. However, for Crizotinib treated cells, the predictions did not overlap with later time-points which might be partly explained by the low resistance evolution in those cells (See [125] for

further details). Furthermore, predicted drug resistance was highest at earlier time-points; at 48h and 3-week, after which decrease was observed.

Similarly, in Erlotinib treated cells we observed a significant increase for cells sampled at day 1 but a slight decrease in later time-points (**Fig. D11b**). This is suggestive of the model capturing a transient, convergent drug-tolerant state rather than the resistance mechanisms. Indeed, quantifying the prediction overlap showed low reconstruction of gene expression $\rho < 0.25$ suggesting that it is non-trivial to apply models trained on bulk data on single-cell data. Furthermore, these observations did not hold for Osimertinib treated samples (**Fig. D11c**) where IC50 predictions did not show any apparent pattern.

5.3 Discussion

Drug resistance is a non-trivial problem in cancer treatment. Multitudes of mechanisms have been described focusing on either individual drugs or individual cancer types albeit with little translational impact. This is partly due to the inherent complexity of biological systems hence the sparsity of repeatable patterns of resistance mechanisms. This issue is further exacerbated by the costly nature of drug discovery and combination studies. Nevertheless, such efforts have led to large-scale datasets to be generated where it is not feasible to capture relevant features with standard relatively low complexity models. For this purpose in a similar spirit with other studies, we hypothesized that large-scale carefully optimized models can delineate common mechanisms/pathways across different tissues and drugs utilizing the relatively vast repository of drug sensitivity measurements. Furthermore, we hypothesized that in order for the extracted features to be relevant in clinical setting, drug sensitivity predictions should be supervised/guided/regularized by patient datasets.

We have generated an integrative model that is aimed at extracting covarying features across patient data and cell-line datasets. Further, we allowed for combining mutation and gene-expression data to aggregate and simplify the multi-layered nature of the cell architecture.

Using the trained model, we showed that the majority of the drug pairs had positive corre-

lations in terms of resistance/sensitivity mechanisms underscoring the need for improved models. Consequently, we further looked into the pairs of drugs that showed higher frequency of possible synergistic and collateral sensitive mechanism and mined for possible common interactors. We identified a *BCL2,HSP90AA1,EGFR* axis with possible synergistic activity and validated in a limited in-vitro setting with *HSP90AA1* and *BCL2* inhibitors.

The utility of the presented model was validated at least in regards to drug Cisplatin where gene-expression features overlapped with previous published signature [141] and features overlapping with previous findings such as *BRAF* mutations sensitizing to ERK-MEK inhibitors. Additionally, drug sensitivity predictions overlapped with progression-free survival data in BLCA, STAD and OV cancers further validating clinical utility.

Furthermore, we investigated whether the trained model was applicable to scRNA-Seq data in a resistance evolution setting. Interestingly, scRNA-Seq data overlapped with the sensitivity predictions in 2 drugs; Alectinib and Erlotinib but failed to associate with transcriptional patterns in Crizotinib treated cells. Furthermore, both Alectinib and Erlotinib treated cells showed an early increase in drug resistance predictions suggesting the prediction of transient resistant states possibly tolerant persister cells.

Nevertheless, the presented results recapitulate the difficulty of mapping the sparse landscape of drug mechanism hence limited utility of clinical translation. In order to better characterize the landscape we plan to expand the current endeavour with high-throughput drug screening efforts.

Chapter 4

Discussion

Chapter 6

DISCUSSION

Biological systems are intrinsically multifaceted and heterogeneous. This inherent heterogeneity arising due to stochasticity leads to intractable diversity governed by evolutionary dynamics. Neoplastic processes are not exempt from these laws of nature resulting in highly diverse unregulated cell populations and allow tumors to evade or develop resistance mechanisms.

Fortunately, evolutionary dynamics enabling cell populations with diverse arsenal of response mechanisms, also allows for convergent mechanisms that are targetable. For instance, one common mechanism of tumorigenesis in non-small cell lung cancer occurs due to an activating somatic mutation in *EGFR* gene where specific Tyrosine Kinase Inhibitors (TKIs) such as Gefitinib and Erlotinib are effective agents for treatment. Unfortunately, cancer eventually recurs frequently due to a secondary hit in gatekeeper mutation T790M which prevents the binding of the drug. Third generation of *EGFR* inhibitors such as Afatinib and Osimertinib can be used in such cases albeit with reduced success rates [152]. This apparent tug of war is common in cancer treatment where improved treatment design by better control of evolutionary dynamics is a must to improve outcomes specifically in high-grade tumors where surgery is not feasible and radiation treatment remains as a palliative option.

However, finding such mechanisms is not always feasible and requires accurate mapping of covarying features across high-dimensional feature space where evolution operates which is a non-trivial task considering the heterogeneity and sparsity of such associations. Towards that goal, complex machine-learning (ML) models even though being 'black-box' have been utilized effectively, specifically recently with the advent of multi-omics single-cell sequencing methods [153, 154].

With a similar spirit, we aimed to mine -omics datasets of relatively large patient cohorts. We showed that such approaches can delineate functional groups of patients both in Acute Myeloid Leukemia (AML) and Myelodysplastic Neoplasms (MDS) in unsupervised fashion (Chapter 2 and Chapter 3). Generated clusters provide a novel view of primary and secondary AML disease prognosis, and clinical relevance of co-mutations in MDS. Furthermore, we also indicate the limitations and strategies to improve data-mining workflows in heterogeneous scRNA-Seq dataset in resistance evolution setting. Finally, we provide a unifying view in terms of manifold learning via multi-omics integration in order to mine for convergent features associated with drug mechanisms (Chapters 4 and 5).

6.1 Model-based clustering of Acute Myeloid Leukemia Patients

We aimed to cluster mutation profiles of AML patients in order to discern clinically relevant co-mutation patterns in a cohort of 2681 patients with complete mutation information for 44 genes. Using Latent Class Analysis (LCA) in a consensus fashion, we clustered the patients/observations in to genomic clusters. We employed the consensus clustering by randomly subsampling the observations and features over 1000 runs. Each subsample is then clustered using LCA and selecting the optimal number of clusters based on silhouette score. Keeping track of co-clustering patterns of observations we generated a consensus matrix which is further clustered using hierarchical clustering to assign the final clusters. We identified 4 genomic-clusters showing distinct clinical profiles. Furthermore, we categorized previously defined primary AML and secondary AML into novel clusters with varying frequency of co-clustering which were not discernible without taking molecular information into account. However, one limitation of the study is that the mutations are aggregated over genes which might prove to be suboptimal.

6.2 Distance-based clustering of Myelodsplastic Neoplasms

In a similar fashion to AML study, we aimed to develop novel genomic/molecular clusters of MDS patients in an unbiased/unsupervised approach. However, we observed increased

heterogeneity in MDS compared to AML which resulted in 'underfitting' when we used model-based approach such as LCA hence we opted to use potentially more sensitive approach based on Autoencoders. Since Autoencoders are efficient frameworks for manifold-learning, we used a single-layer network to dimension reduce the binary mutation profiles of MDS patients. This is similar to binary Principal Component Analysis or Multiple Correspondence Analysis for categorical data however allowing for non-linear associations to be mapped as well. Using cross-validation to select for the number of dimensions to embedd, we applied this strategy in a consensus fashion and identified 14 molecular clusters. Identified clusters showed high clinical relevance in terms of overall survival and treatment response consequently resulting in unbiased, molecularly defined subgroups. Exclusion of clinical parameters however, although simplifying, is possibly suboptimal. Furthermore, distance based models, given large enough observations, will potentially stratify all cases into unique combinations of binary mutations hence future efforts to regularize the stratification process via supervision with unbiased clinical parameters can potentially benefit the current molecular clusters.

6.3 Benchmarking scRNA-Seq analysis workflows

Transitioning from leukemia to solid tumors, we focused on the utility of available tools for scRNA-Seq analysis. Since it is not possible to fully capture the heterogeneous nature of tumors via bulk sequencing, single-cell approaches have become the gold standard. However, it is non-trivial to use single-cell rna-sequencing due to sparsity and increased dropout rates. In order to guide current research, we conducted a comprehensive benchmarking analysis in relatively heterogeneous setting of resistance evolution. We showed how prone the overall analysis in terms of clustering and trajectory mapping to parameter and method selection. Specifically, non-linear dimension reduction methods lead to qualitative results difficult to quantitative investigations where regularization in the form of priors can alleviate such issues. We further evaluated the utility of supervision in identifying transcriptional

dynamics which are of interest in order to capture regulatory mechanisms, specifically in drug resistance settings. Our results can further be extended to multi-modal single-cell datasets as well where simultaneous profiling of mRNA, gDNA, chromatin accessibility can be performed.

6.4 Integrative Modeling of Drug Sensitivities

Finally as a proof of concept study, we investigated the utility of using neural-networks in manifold-learning objective regularized by drug sensitivity measurements effectively modeling the evolutionary landscape. We used publicly available cell-line and patient datasets from GDSC, DepMap and TCGA repositories. Novelty of our approach is based on the simultaneous optimization of patient data, cell-line data and drug sensitivities effectively characterizing the fitness landscape associated with drug responses. Since predictive models, specifically highly complex models, are difficult to generalize to patient datasets, we aimed to identify features predictive of drug sensitivity relevant to both cell-lines and patient samples. Further applying the proposed method in a tissue specific manner, we were able to draw associations between bulk RNA-Seq and scRNA-Seq under resistance evolution setting. We aim to further extend this study and expand to in-vitro validation by large mono and in combination screening.

6.5 Conclusion

As our understanding of the inherent complexity of biological systems specifically in the temporal setting grows, we are driven towards heuristics to alleviate the *curse of dimensionality* inherent in -omic datasets. However, clinical translation of cancer research and/or better characterization of cancer progression and drug response requires models that can integrate multi-modal feature space. For this purpose machine-learning approaches, whether in black-box nature or in the dimension reduction settings, can provide powerful frameworks for extracting features relevant to the phenotype of interest. In order to investigate the utility of such approaches in diverse settings, we developed models i) applicable in clinically rele-

vant stratification of myeloid neoplasms, ii) that can integrate multi-omic features. Finally, we have also studied the variation inherent in single-cell analysis workflows underscoring the need for robust unbiased approaches for reproducibility in research. We hope that such strategies will enable clinical translation of cancer research in an unbiased and reproducible fashion.

Appendix A

MACHINE LEARNING INTEGRATES GENOMIC SIGNATURES FOR SUBCLASSIFICATION BEYOND PRIMARY AND SECONDARY ACUTE MYELOID LEUKEMIA

A.0.1 Supplementary Tables

Table A1: **Summary of all sources of Acute Myeloid Leukemia (AML) cases included in our study**

Cohorts	**Total** (n=6991)
Our group (n=4857)	
Munich Leukemia Laboratory (MLL)	4002
Cleveland Clinic (CC)	855
Publicly Available AML cohorts (n=1931)	
The Cancer Genome Atlas (TCGA)	182
German-Austrian Study Group	1251
Beat AML Master Trial	498
External validation cohort	
The University of Texas MDS Anderson Cancer Center (MDACC)	203

Table A2: **List of 44 genes on the targeted sequencing panel**

ASXL1	CEBPA	ETV6	IDH2	KRAS	RAS	RUNX1	SRSF2	TP53
BCOR	CSF1R	EZH2	JAK2	LUC7L2	PHF6	SETBP1	STAG2	U2AF1
BCORL1	CUX1	FLT3	KDM6A	MECOM	PRPF8	SF3B1	STAT3	WT1
CALR	DNMT3A	GATA2	KIT	NF1	PTPN11	SIMC1	SUZ12	ZRSR2
CBL	EED	IDH1	KMT2A	NPM1	RAD21	SMC3	TET2	

Table A3: **Gene mutation frequencies in Acute Myeloid Leukemia by Subtype**

Mutant Gene	CBF-AML n (%)	APL n (%)	KMT2A-AML n (%)	pAML n (%)	sAML n (%)	tAML n (%)
ASXL1	24 (6)	8 (3.8)	36 (11.5)	461 (12.7)	152 (24.4)	18 (7.9)
BCOR/BCORL1	7 (1.8)	1 (0.7)	8 (3.5)	162 (6.2)	49 (9.0)	3 (2.1)
CBL	15 (3.8)	4 (1.8)	5 (1.8)	71 (2.6)	26 (4.5)	9 (5.0)
$CEBPA^{Mo/Bi}$	3 (0.9)	7 (4.2)	0 (0)	390 (12.0)	25 (4.2)	11 (5.2)
DNMT3A	15 (3.8)	3 (1.4)	54 (18.6)	1089 (31.5)	81 (14.0)	35 (18.2)
ETV6	11 (2.9)	7 (3.3)	3 (1.3)	59 (2.2)	20 (3.7)	2 (1.4)
EZH2	13 (3.4)	4 (2.0)	5 (2.2)	105 (4.2)	32 (5.8)	5 (3.4)
$FLT3^{ITD/TKD}$	67 (17.7)	138 (55.8)	63 (25.3)	1101 (36.4)	83 (14.4)	60 (33.3)
GATA2	4 (1.4)	5 (2.5)	6 (2.6)	115 (4.3)	14 (2.5)	2 (1.4)
IDH1	4 (1.1)	1 (0.5)	17 (6.2)	324 (12.1)	38 (6.9)	16 (8.6)
IDH2	7 (1.8)	1 (0.5)	39 (13.8)	482 (17.8)	54 (10.0)	17 (9.9)
KIT	90 (22.2)	2 (1)	3 (1.3)	46 (1.8)	10 (1.8)	1 (0.6)
KRAS	39 (10.1)	9 (4.1)	35 (12.0)	151 (4.8)	23 (4.0)	19 (9.6)
NPM1	6 (2.2)	0 (0)	8 (3.5)	958 (35.9)	32 (6.1)	23 (15.2)
NRAS	119 (33.9)	9 (4.1)	49 (17.9)	421 (15.6)	57 (10.5)	28 (15.4)
RUNX1	7 (1.8)	5 (2.3)	40 (12.0)	642 (16.9)	179 (26.0)	31 (11.6)
SF3B1	1 (0.3)	2 (1.0)	5 (2.0)	127 (4.5)	61 (10.5)	18 (4.4)
SRSF2	6 (1.6)	3 (1.5)	21 (9.3)	334 (13.1)	212 (22.1)	13 (8.8)
STAG2	2 (0.1)	0 (0)	19 (9.0)	110 (5.4)	55 (10.8)	9 (6.9)
TET2	44 (11.1)	4 (1.9)	24 (8.3)	639 (20.0)	114 (19.8)	25 (14.1)
TP53	5 (1.3)	3 (2.2)	57 (18.2)	288 (8.6)	71 (12.0)	46 (20.4)
U2AF1	3 (0.7)	0 (0)	11 (4.8)	102 (4.1)	47 (8.6)	4 (2.8)
WT1	21 (5.5)	38 (17.9)	12 (4.1)	262 (7.4)	23 (3.7)	8 (3.6)
ZRSR2	1 (0.3)	5 (2.5)	4 (1.8)	49 (2.0)	16 (2.9)	1 (0.7)

Table A4: **Multivariate Cox-Proportional Hazards model determines genomic features associated with survival and disase risk 'Favorable vs Adverse' in various acute myeloid leukemia cohorts.**

AML Cohort	Favorable	Adverse
All (n=6788)	t(15;17)*, inv(16)/t(8;21)*, CEBPABi *, NPM1*	-5/del(5q), inv(3)/t(3;3), -17/del(17p), ASXL1*, DNMT3A*, KRAS, FLT3ITD*, RUNX1, SRSF2*, TET2, TP53*, U2AF1
KMT2AR-AML (n=371)		del(9p), -Y, KRAS, GATA2, TP53
APL (n=312)		ASXL1, SF3B1
CBF-AML (n=422)		SRSF2, U2AF1
inv(16)/t(16;16) (n=230)		KRAS, FLT3, U2AF1
t(8,21) (n=192)		RUNX1, U2AF1
NK-AML (n=3288)	NPM1	ASXL1, DNMT3A*, FLT3ITD*, SRSF2*, U2AF1
AK-NCK-AML (n=1695)		inv(3)/t(3;3), -5/del(5q), del(16q), ASXL1, DNMT3A, EZH2, KRAS, SRSF2, TET2, TP53, U2AF1*
CK-AML (n=702)	GATA2, KIT	t(1;22), CBL, DNMT3A, EZH2, KRAS*, RUNX1, TP53*
AML with specific cytogenetic abnormalities		
inv(3)/t(3;3) (n=80)	SF3B1	-5/del(5q), KRAS, DNMT3A
-5/del(5q) (n=458)	BCOR/L1	del(9p), KRAS, RUNX1, SRSF2, TP53*, U2AF1*
-7/del(7q) (n=461)	IDH1	t(1;22), del(9p), DNMT3A, KIT, KRAS, NRAS, TP53*, U2AF1, WT1
del(12p) (n=150)	del(20q), ASXL1	-5/del(5q), del(13q), U2AF1
-17/del(17p) (n=202)		del(13q), ETV6, IDH1, KRAS, TP53, U2AF1
del(20q) (n=109)	-5/del(5q)	del(13q), GATA2, IDH1, TET2
+8 (n=562)	ZRSR2	inv(3)/t(3;3), -6/del(6q), EZH2, KRAS, RUNX1, SRSF2, TP53, U2AF1
AML with specific gene mutations		
ASXL1 (n=631)		GATA2, NRAS
BCOR/L1 (n=214)		-6/del(6q), JAK2
CBL (n=106)		-5/del(5q), 17/del(17p), -Y, ASXL1, ETV6, IDH1
CEBPA (n=426)	BCOR/L1	-5/del(5q), del(12p), del(16q), +8, NPM1, RUNX1, SRSF2, ZRSR2
$CEBPA^{Mo}$ (n=236)	BCOR/L1, ETV6	del(16q), RUNX1, SF3B1, TET2, ZRSR2
$CEBPA^{Bi}$ (n = 190)		NPM1
DNMT3A (n=1205)		inv(3)/t(3;3), FLT3ITD, JAK2, SRSF2, TP53, U2AF1
ETV6 (n=81)		FLT3, GATA2, SRSF2, U2AF1, WT1
EZH2 (n=142)	IDH1, IDH2, WT1	-5/del(5q), del(12p), +8, NRAS, RUNX1, SRSF2, U2AF1
FLT3ITD/TKD (n=1244)	NRAS	DNMT3A, ETV6, KIT, RUNX1, TET2, ZRSR2
GATA2 (n=131)	-7/del(7q), -17/del(17p)	ASXL1, BCOR/L1, ETV6, JAK2, KRAS, NPM1, RUNX1*, TET2, TP53
IDH1 (n=378)	-7/del(7q), NPM1, NRAS	-17/del(17p), del(20q), DNMT3A, JAK2
IDH2 (n=553)	NPM1*	DNMT3A, ETV6, JAK2, SRSF2
KIT (n=57)	+8, CBL, SF3B1	-7/del(7q), ASXL1, FLT3ITD, TP53, U2AF1
KRAS (n=193)	FLT3TKD	-5/del(5q), RUNX1, TP53
NPM1 (n=1013)	NRAS	DNTM3A, ETV6, FLT3
NRAS (n=506)	NPM1*	-6/del(6q), -X, TET2
RUNX1 (n=852)	BCOR/L1	DNMT3A, JAK2*, SRSF2
SF3B1 (n=206)		del(13q), DNMT3A
SRSF2 (n=559)	IDH2, NPM1	ETV6, KIT
Cohesion complex (n=285)	KDM6A	inv(3)/t(3;3), FLT3ITD, NRAS, SRSF2, TP53
TET2 (n=778)		-6/del(6q), FLT3ITD, KRAS, NRAS, U2AF1
TP53 (n=405)		-5/del(5q), -6/del(6q)
U2AF1 (n=153)		-5/del(5q), EZH2, KRAS
WT1 (n=293)	TET2	-7/del(7q)*, -X, -Y, DNMT3A*, SRSF2*
ZRSR2 (n=66)		ASXL1, ETV6

Table A5: **Comparison of baseline and clinical characteristics of primary versus secondary acute myeloid leukemia**

Variables	pAML.n.(%)	sAML.n.(%)	P-value
Age* (y), median (range)	66.9 (18-89)	70 (21-89)	< 0.0001
≥ 60 y	2374 (65.5)	638 (81.3)	< 0.0001
Gender			
Male	2373 (52.7)	535 (64.4)	< 0.0001
Female	2129 (47.3)	297 (35.6)	
Hematological Parameters*			
WBC (109/L), median (range)	20.2 (0.1-600)	5.3 (0.5-388)	< 0.0001
< 3 x 109/L	874 (20.4)	279 (36.6)	< 0.0001
Hemoglobin (g/dL), median (range)	9.2 (2.3-17.9)	9.3 (5-16.5)	0.3
< 10 g/dL	2479 (65.9)	484 (66.3)	0.5
Platelets (109/L), median (range)	73 (2-2366)	50 (5-869)	<0.0001
< 100 x 109/L	2663 (60.6)	573 (76.5)	<0.0001
Bone marrow			
Blasts %, median (IQR)	61 (48)	30 (36)	<0.0001

Table A6: **Probability of survival per each genomic cluster**

Survival	**Cluster-1**	**Cluster-2**	**Cluster-3**	**Cluster-4**
	Probability (% Confidence Interval)			
1-year	0.69 (0.65-0.72)	0.71 (0.68-0.74)	0.59 (0.55-0.63)	0.41 (0.35-0.48)
2-year	0.55 (0.51-0.59)	0.52 (0.48-0.55)	0.35 (0.31-0.40)	0.23 (0.18-0.30)
3-year	0.50 (0.46-0.54)	0.44 (0.41-0.48)	0.27 (0.23-0.32)	0.17 (0.12-0.24)
5-year	0.43 (0.39-0.47)	0.37 (0.33-0.40)	0.19 (0.15-0.23)	0.13 (0.08-0.20)

Table A7: **Characteristics of the secondary acute myeloid leukemia cases in genomic cluster-1**

AML subtype	**Antecedent diagnosis**	**Age (y)**	**Sex**	**OS (months)**	**Cytogenetics**	**Somatic Mutations**
sAML		48	F	78.50	del(9)(q13q22)	NPM1, NRAS
sAML	MDS	85	M	7.90	Normal	NPM1
sAML	MDS	62	F	5.30	t(3;12)(q26.2;p13)	NPM1
sAML	MDS	76	M	13.13	-Y	FLT3TKD, IDH1, NPM1, SRSF2
sAML	MDS	65	M	9.60	Normal	IDH2, NPM1, SRSF2
sAML	MDS	63	F	10.63	Normal	IDH2, NPM1
sAML	MDS	81	M	4.17	Normal	IDH2, NPM1, NRAS
sAML	MDS	52	F	20.57	-X	NPM1
sAML	MDS	60	F	8.63	8	NPM1, NRAS
sAML			M	55.17	Normal	DNMT3A, NPM1, TET2, WT1
sAML			F	6.60	Normal	NPM1, NRAS
sAML			F	0.87	Normal	CEBPAMo, NPM1
sAML			F	2.43	Normal	IDH2, NPM1, SRSF2
sAML			F	10.77	Normal	DNMT3A, FLT3ITD, NPM1
sAML			M	67.77	Normal	FLT3ITD, NPM1
sAML			F	10.93	Normal	DNMT3A, NPM1
sAML			F	60.73	Normal	DNMT3A, NPM1
sAML	MDS	74	M	8.63	del(7)(q22q36)	NPM1
sAML	MDS	87	M	6.40	Normal	BCOR, NPM1, SF3B1, TET2
sAML	MDS	68	F	11.30	Complex	ETV6, NRAS, NPM1
sAML	MDS	62	M	45.50	t(3;11))(p21;q23)	NPM1
sAML	MDS	84	M	24.3	Complex	NPM1

Table A8: **Baseline and clinical characteristics of patients in each genomic cluster**

Variables	**Cluster-1** n (%)	**Cluster-2** n (%)	**Cluster-3** n (%)	**Cluster-4** n (%)
Age (y), median (95% CI)	62 (60 - 63.8)	60 (58.3 - 61.4)	70 (69 - 71.1)	70 (67.8 - 71.8)
Hematological Parameters				
WBC (109/L), median (95% CI)	29.8 (25.2 - 34.8)	9.5 (8 - 11.9)	8.9 (7.1- 11.3)	4.4 (3.5 - 5.2)
< 3 x 109/L, freq (95% CI)	0.1 (0.09 - 0.14)	0.28 (0.25 - 0.31)	0.28 (0.24 - 0.31)	0.36 (0.31 - 0.42)
Hgb (g/dL), median (95% CI)	9.1 (9 - 9.3)	9.3 (9.1 - 9.4)	9 (8.9 - 9.3)	9 (8.6 - 9.2)
< 10 g/dL, freq (95% CI)	0.66 (0.63 - 0.7)	0.62 (0.59 - 0.65)	0.67 (0.63 - 0.7)	0.75 (0.7 - 0.79)
Plt (109/L), median (95% CI)	60 (55-64)	56 (52-61)	59 (55-65)	51 (41-58)
< 100 x 109/L, freq (95% CI)	0.7 (0.67 - 0.73)	0.74 (0.71 - 0.77)	0.73 (0.69 - 0.76)	0.8 (0.74 - 0.84)
Bone Marrow				
Blasts %, median (95% CI)	77 (73-80)	60 (54-64)	52 (50-56)	41 (36-48)

Table A9: **Baseline and clinical characteristics of primary and secondary acute myeloid leukemia cases from the MD Anderson Cancer Center cohort**

Variables	**pAML** (n=143)	**sAML** (n=60)
Age, (y), median (range)	69 (24-86)	69 (28-83)
F/M	75/68	21/39
Hematological parameters		
WBC (109/L), median (range)	3.6 (0.6-143.7)	3.4 (0.3-85.6)
Hgb (g/dL), median (range)	9.2 (3.9-11.6)	8.7 (6.4-10.7)
Plt (109/L), median (range)	40 (7-271)	38 (2-864)
Bone marrow		
Blasts %, median (range)	14 (0-97)	11 (0-70)

Table A10: **Frequencies of cytogenetic abnormalities and gene mutations in primary and secondary acute myeloid leukemia cases from the MD Anderson Cancer Center cohort**

Variable	**pAML** (%)	**sAML** (%)	**Total** (%)
Cytogenetic abnormalities			
inv(3)/t(3;3)	2.10	6.67	3.45
t(6;9)	2.10	0.00	1.48
-5/del(5q)	20.98	25.00	22.17
del6q/-6	6.29	5.00	5.91
-7/del(7q)	16.08	33.33	21.18
-9/del(6q)	4.20	3.33	3.94
del(12p)	0.00	0.00	0.00
del(13q)	0.00	0.00	0.00
del(16q)	0.00	0.00	0.00
-17/del(17p)	11.89	10.00	11.33
del(20q)	1.40	3.33	1.97
8	11.89	16.67	13.30
-X	4.90	3.33	4.43
-Y	1.40	3.33	1.97
Gene mutations			
ASXL1	20.28	43.33	27.09
BCOR/L1	20.28	28.33	22.66
CBL	3.50	5.00	3.94
CEBPAMo	7.69	10.00	8.37
CEBPABi	0.70	0.00	0.49
DNMT3A	28.67	18.33	25.62
ETV6	0.70	10.00	3.45
EZH2	15.38	21.67	17.24
FLT3TKD	30.07	5.00	22.66
FLT3ITD	25.17	3.33	18.72
GATA2	6.29	10.00	7.39
IDH1	11.19	3.33	8.87
IDH2R140	11.19	3.33	8.87
IDH2R172	1.40	3.33	1.97
KIT	1.40	0.00	0.99
KRAS	6.99	8.33	7.39
NPM1	13.29	26.67	17.24
NRAS	27.27	0.00	19.21
RUNX1	19.58	25.00	21.18
SF3B1	3.50	8.33	4.93
SRSF2	16.08	20.00	17.24
TET2	42.66	41.67	42.36
TP53	27.97	31.67	29.06
U2AF1	4.20	21.67	9.36
WT1	4.90	1.67	3.94
ZRSR2	4.90	8.33	5.91

A.0.2 Supplementary Figures

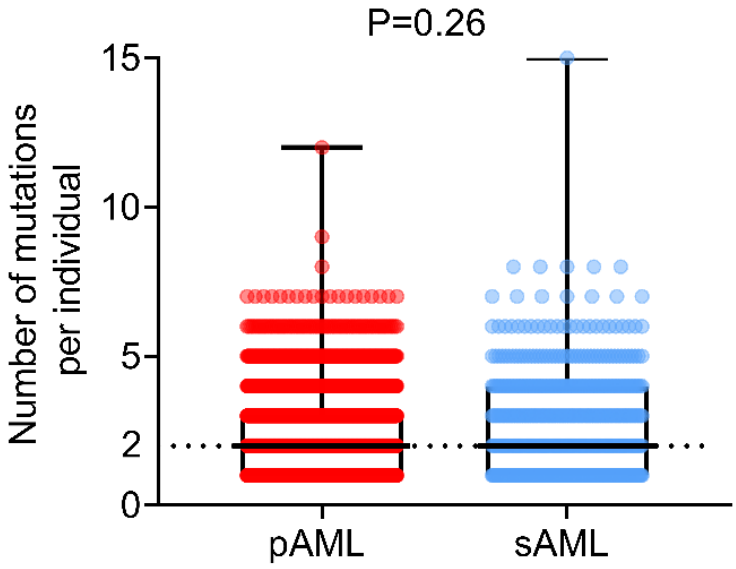


Figure A1: Comparison of the mutational burden in acute myeloid leukemia subtypes. The plot represents number of somatic mutations per individuals in primary (pAML) vs. secondary acute myeloid leukemia (sAML). Levels of statistical significance is indicated using p-value.

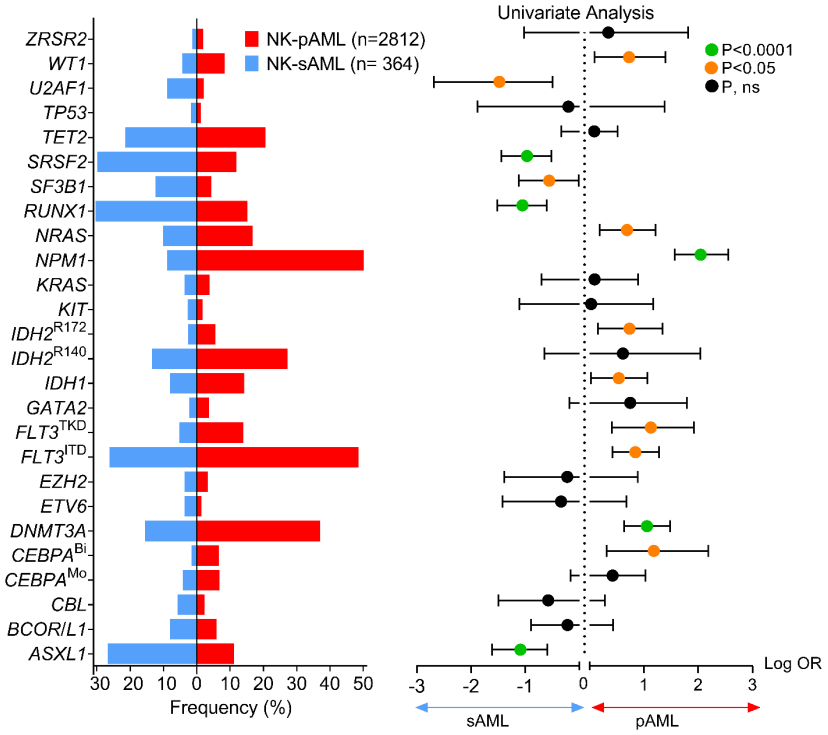


Figure A2: Comparison of somatic mutations associated with abnormal normal karyotype primary versus secondary acute myeloid leukemia. A bar graph showing the frequency (in percent) of somatic mutations in normal karyotype primary (NK-pAML) vs. secondary acute myeloid leukemia (NK-sAML). Forest plots representing univariate analyses showing the odds ratio (OR) of the association of somatic mutations in NK-pAML vs. NK-sAML. Levels of statistical significance are indicated in green, orange, and black colors (P<0.0001, P<0.05, and P>0.05, respectively) using fisher's exact test. The abbreviation ns denotes non-significant.

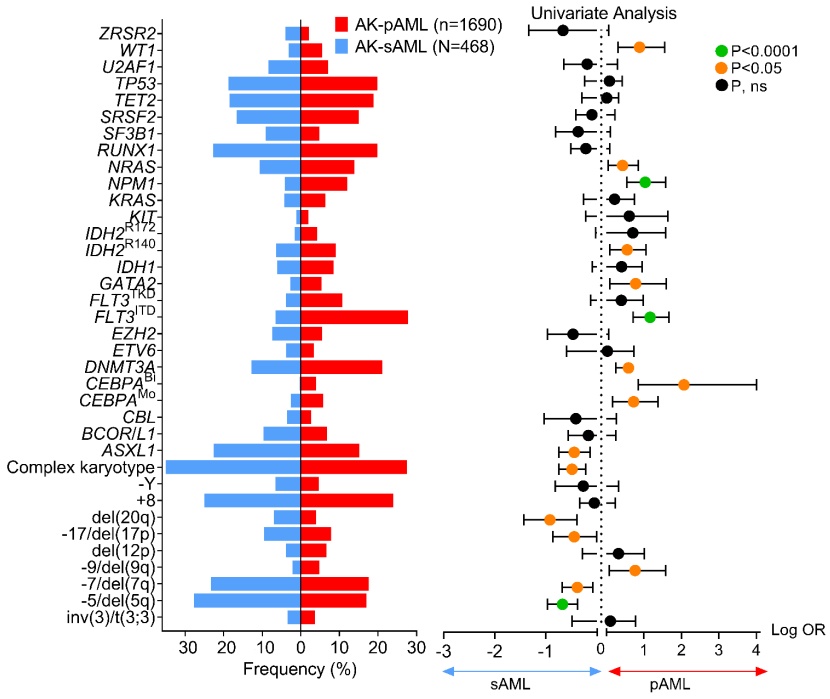


Figure A3: Comparison of somatic mutations and cytogenetic abnormalities associated with abnormal karyotype primary versus secondary acute myeloid leukemia. A bar graph showing the frequency (in percent) of somatic mutations and cytogenetic abnormalities in abnormal karyotype primary (AK-pAML) vs. secondary acute myeloid leukemia (AK-sAML). Forest plots representing univariate analyses showing the odds ratio (OR) of the association of somatic mutations in AK-pAML vs. AK-sAML. Levels of statistical significance are indicated in green, orange, and black colors (P<0.0001, P<0.05, and P>0.05, respectively) using fisher's exact test. The abbreviation ns denotes non-significant.

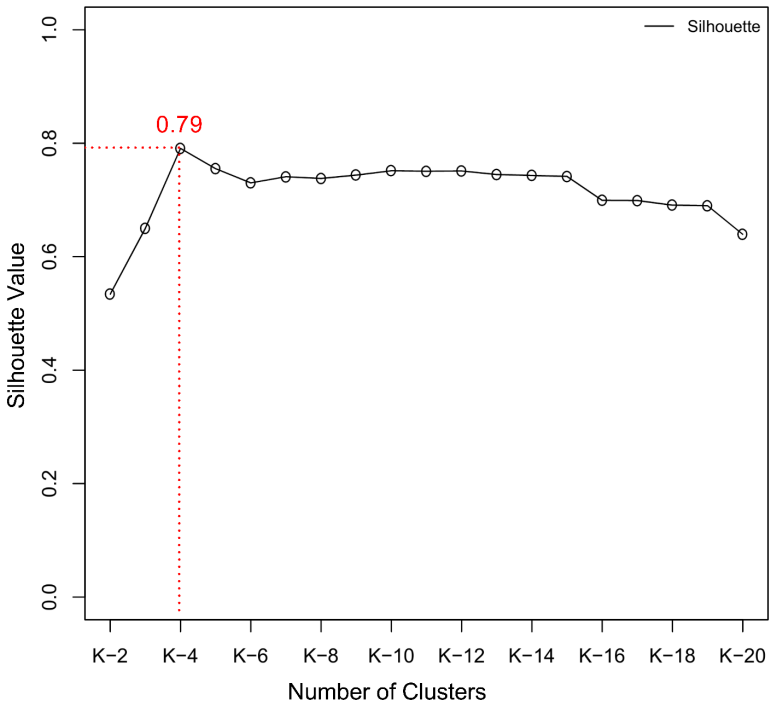


Figure A4: Silhouette value and selection of number of genomic clusters. The plot represents the silhouette value with respect to the number of clusters that can be identified by Bayesian latent class analysis. As seen, a number of 4 clusters attributes to the highest silhouette value of 0.79. Therefore, we selected 4 clusters based on the silhouette value.

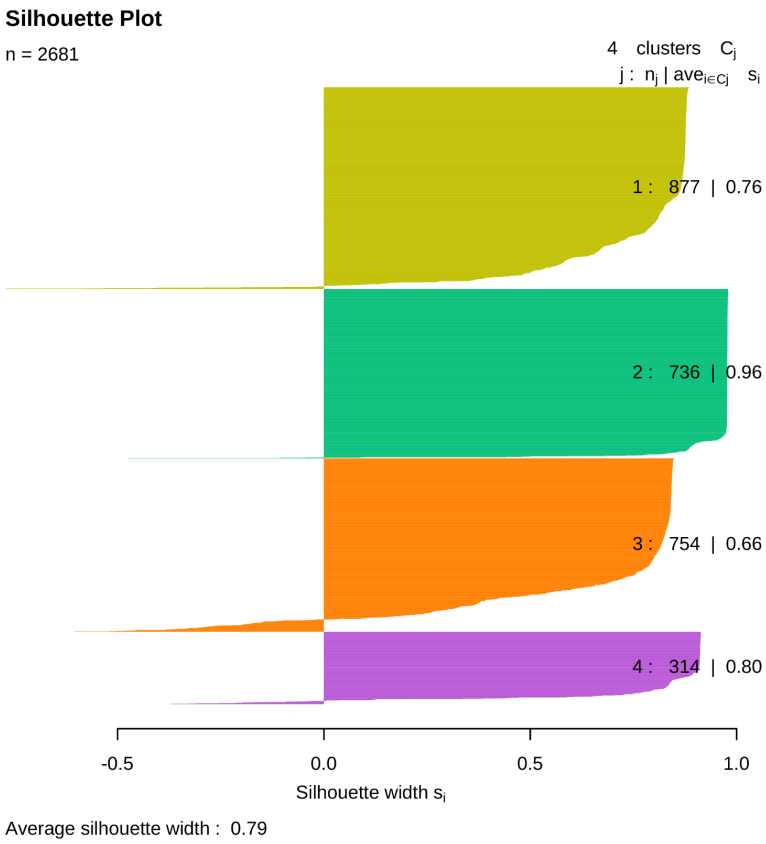


Figure A5: Silhouette value in each genomic cluster. The plot represents the silhouette values in each of the identified clusters. Genomic cluster-1 in yellow, genomic cluster-2 in green, genomic cluster-3 in orange and genomic cluster-4 in purple.

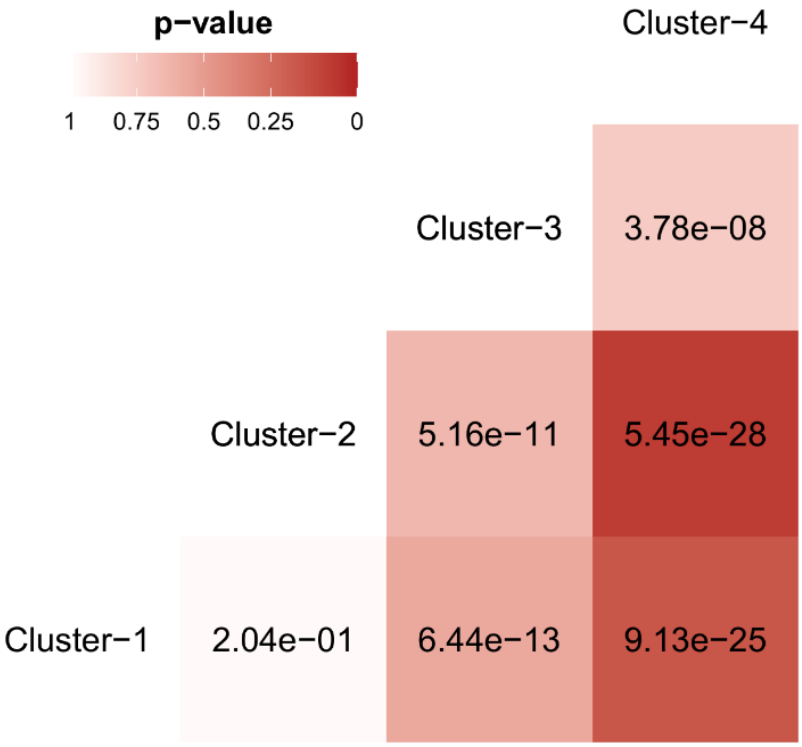


Figure A6: Pairwise survival comparison between the identified genomic clusters. The figure illustrates the pairwise survival tests implemented to assess for the level of significant survival difference between each of the identified genomic clusters (GC).

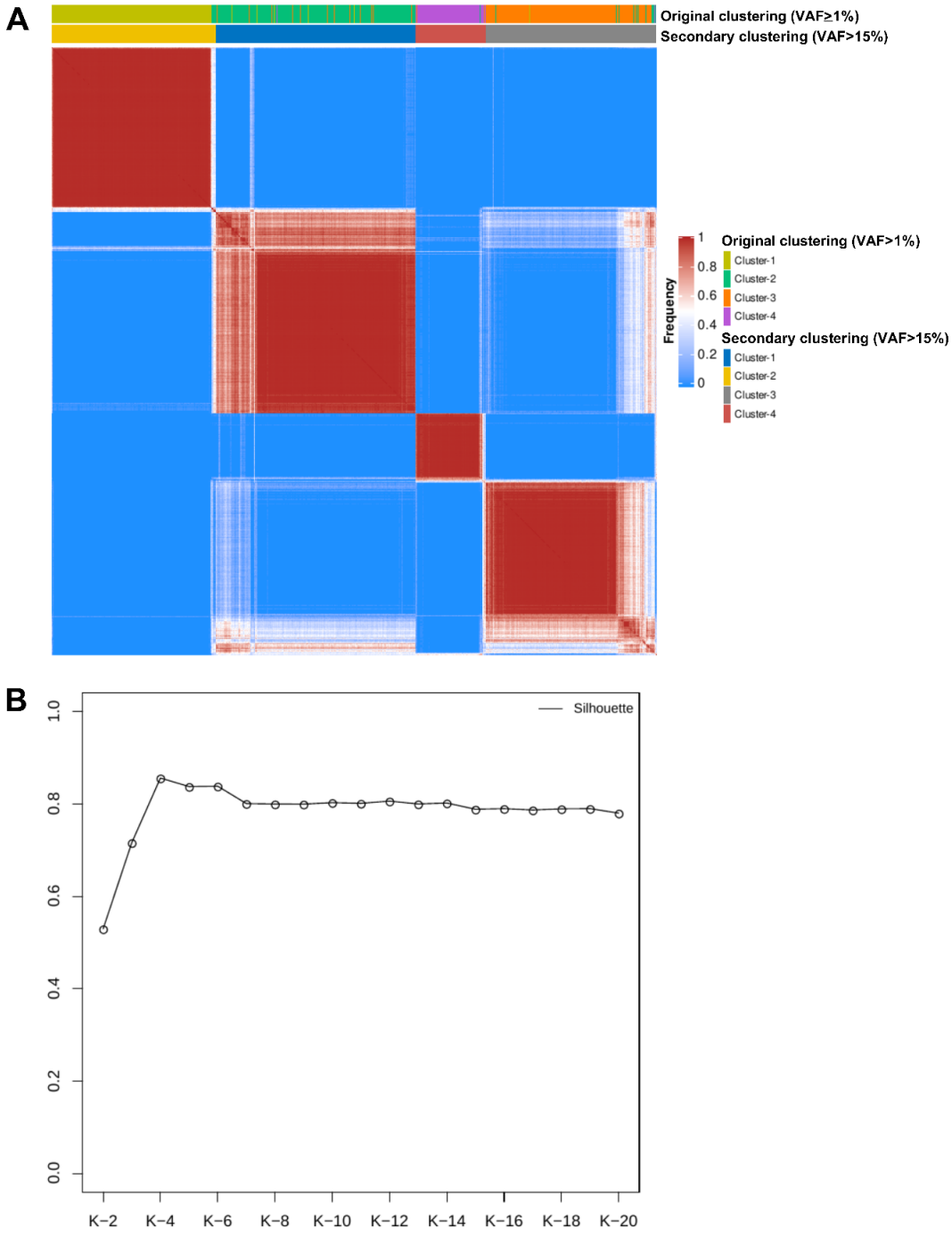


Figure A7: . Results of the Bayesian Latent Class clustering based on the silhouette value when 15% variant allele frequency cut-off is considered. (A) Consensus matrix generated by applying latent class analysis on 1000 subsamples representing the frequency of two observations being clustered in the same group. (B) The plot represents the silhouette value with respect to the number of clusters that can be identified by Bayesian latent class analysis. As seen, a number of 4 clusters attributes to the highest silhouette value of 0.86. Therefore, we selected 4 clusters based on the silhouette value.

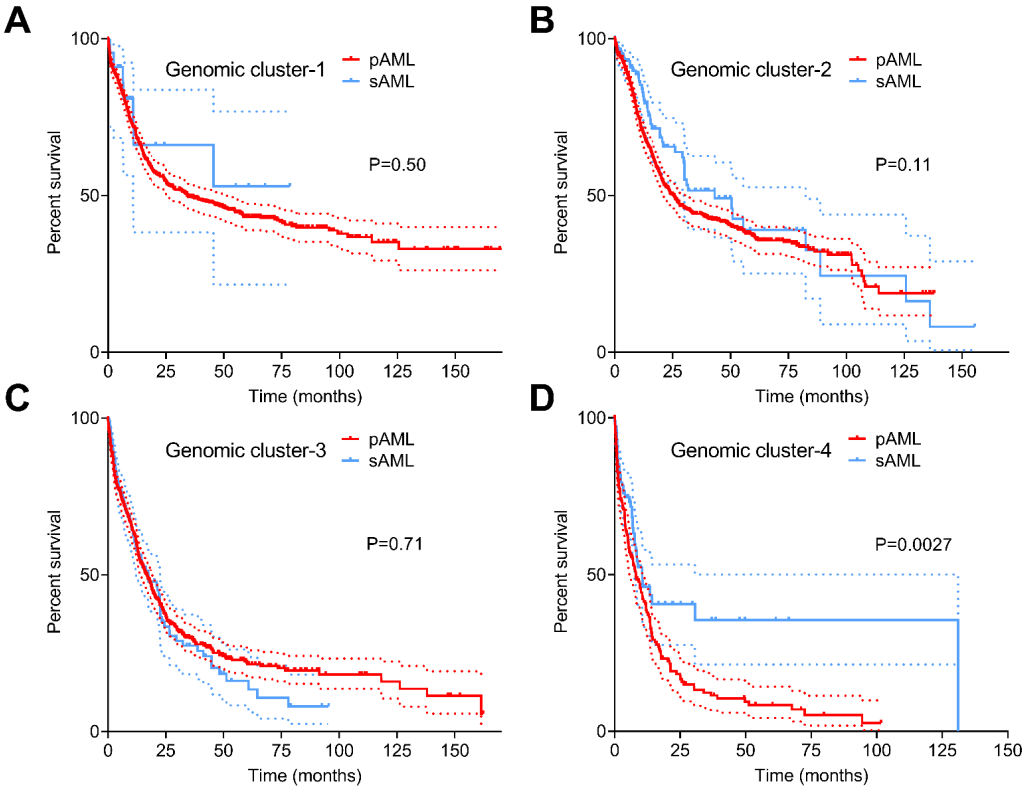


Figure A8: Results of the overall survival comparison of primary versus secondary acute myeloid leukemia within each genomic cluster. (A-D) Kaplan-Meier analyses showing overall survival (in months) of primary vs. secondary acute myeloid leukemia within each cluster. Levels of statistical significance are indicated using p-values.

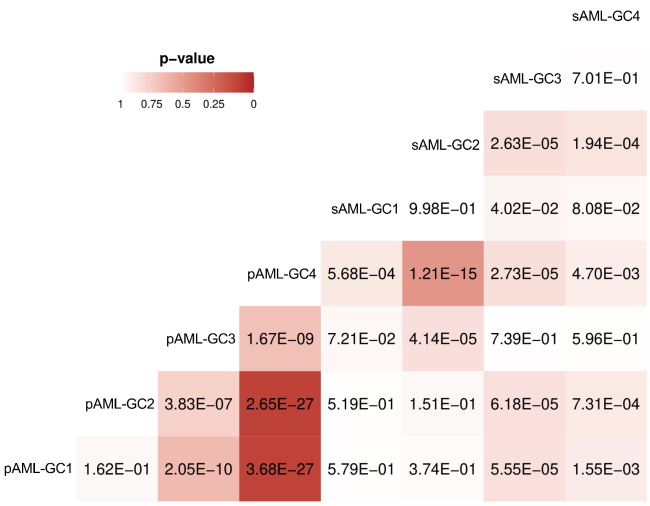


Figure A9: Pairwise survival comparison between acute myeloid leukemia subtypes within each genomic cluster. The figure illustrates the pairwise survival tests implemented to assess for the level of significant survival difference between primary (pAML) and secondary (sAML) acute myeloid leukemia in each of the identified clusters (C; example, C-1 means Cluster-1, etc). Levels of statistical significance are indicated.

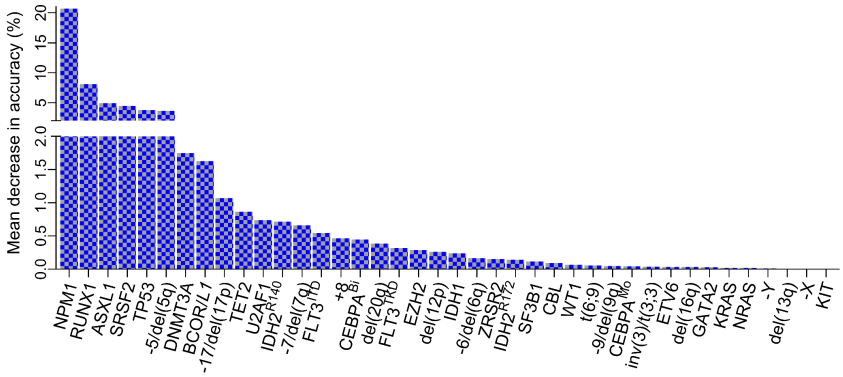


Figure A10: The global importance of genomic signatures in the model. A bar plot showing the genomic features used in our model and their respective importance calculated by mean decrease in accuracy. The y-axis shows the decrease in overall classification accuracy if the given variable is removed from the model.

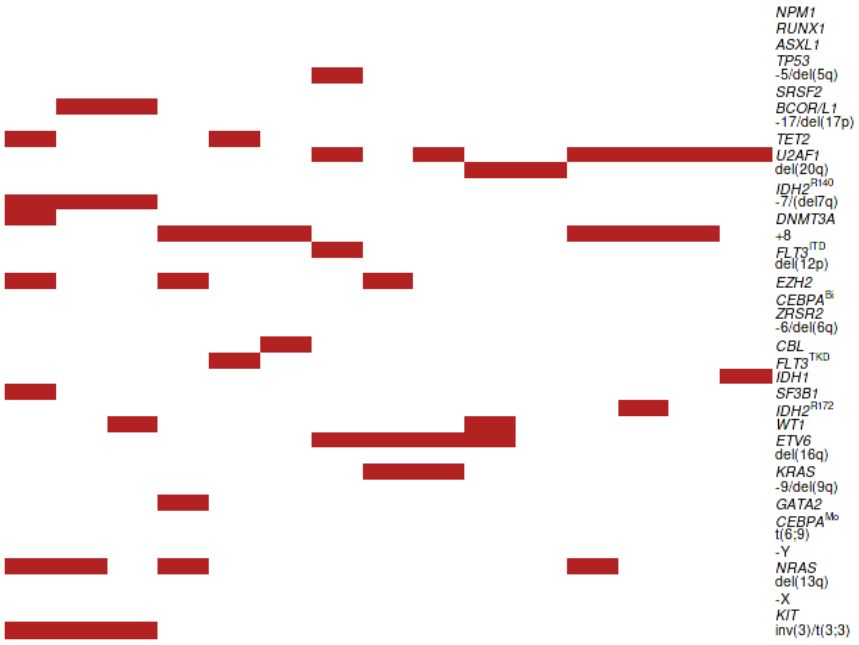


Figure A11: Genomic features characterizing the misclassified cases in genomic cluster 3. A heatmap showing the genomic features of the misclassified cases in genomic cluster 3.

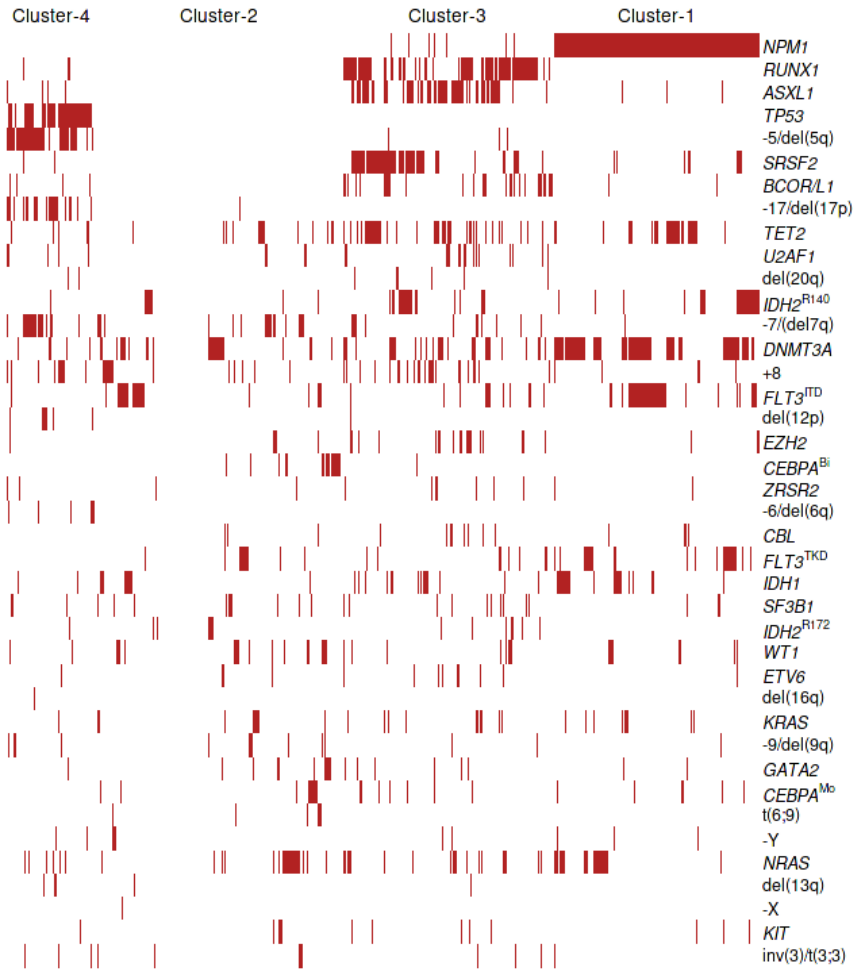


Figure A12: A summary of the invariant genomic features defining each genomic cluster. A heatmap demonstrating the genomic features of each genomic cluster.

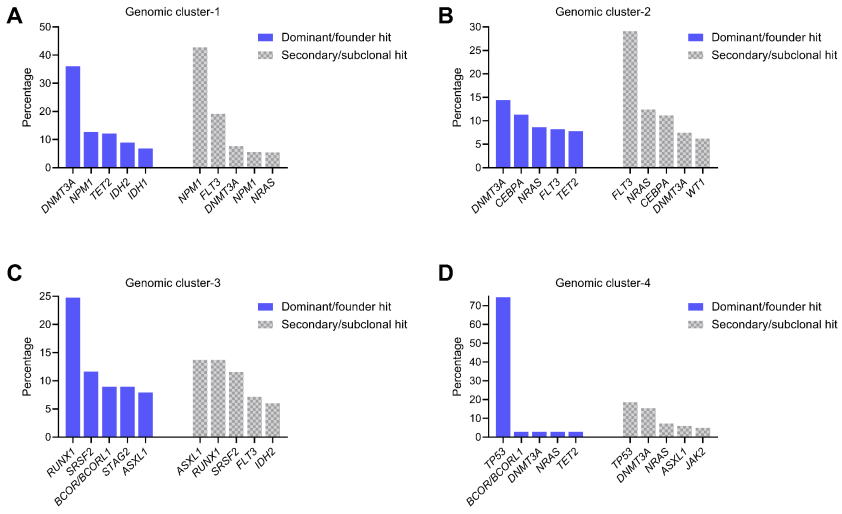


Figure A13: The clonal hierarchy of gene mutations per genomic clusters. The bar graphs represent the top 5 most frequent dominant/founder and secondary/subclonal gene mutations per each genomic cluster (Panel A: genomic cluster-1, Panel B: genomic cluster-2, Panel C: genomic cluster-3 and Panel D: genomic cluster-4) as represented in the figure.

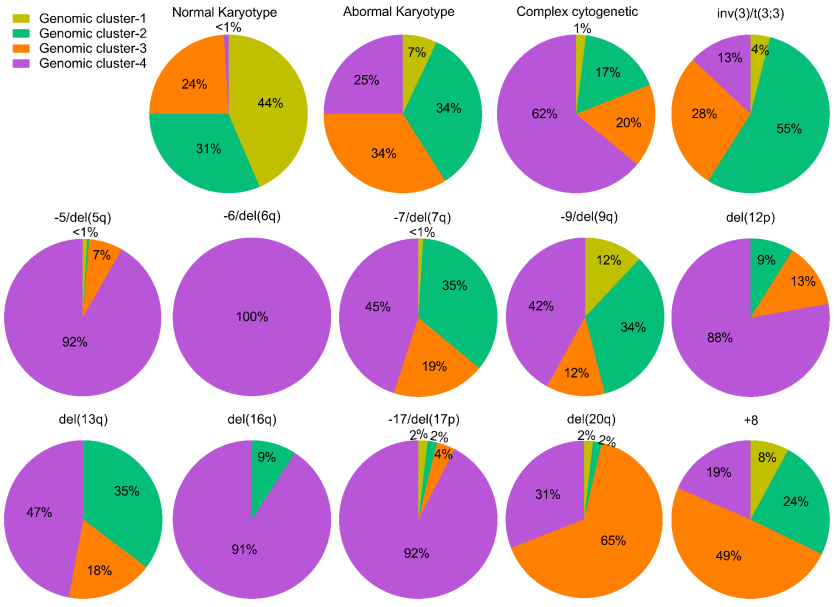


Figure A14: . Genomic clusters' percentages in common cytogenetic abnormalities in acute myeloid leukemia. The pie charts illustrates the percentage of each genomic cluster in several common cytogenetic abnormalities. The figure legends colors are assigned specifically for each genomic cluster.

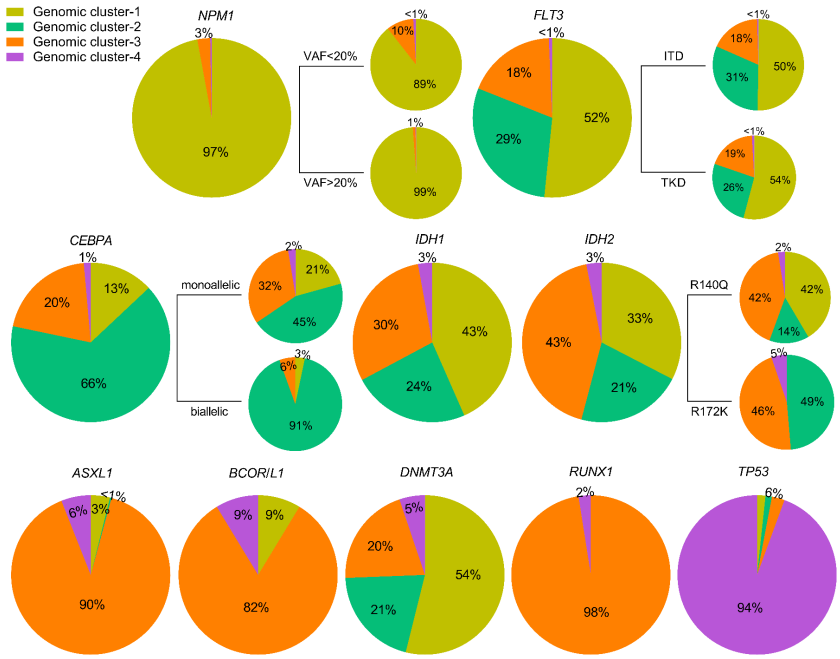


Figure A15: . Genomic clusters' percentages in selected gene mutations in acute myeloid leukemia. The pie charts illustrates the percentage of each genomic cluster in several common gene mutations. The figure legends colors are assigned specifically for each genomic cluster.

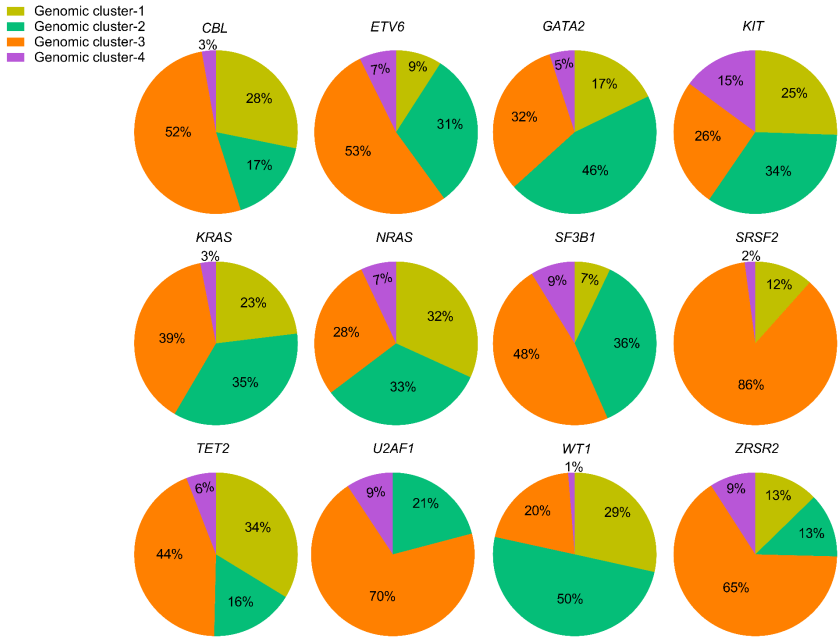


Figure A16: Genomic clusters' percentages in selected gene mutations in acute myeloid leukemia. The pie charts illustrates the percentage each genomic cluster in several common gene mutations. The figure legends colors are assigned specifically for each genomic cluster.

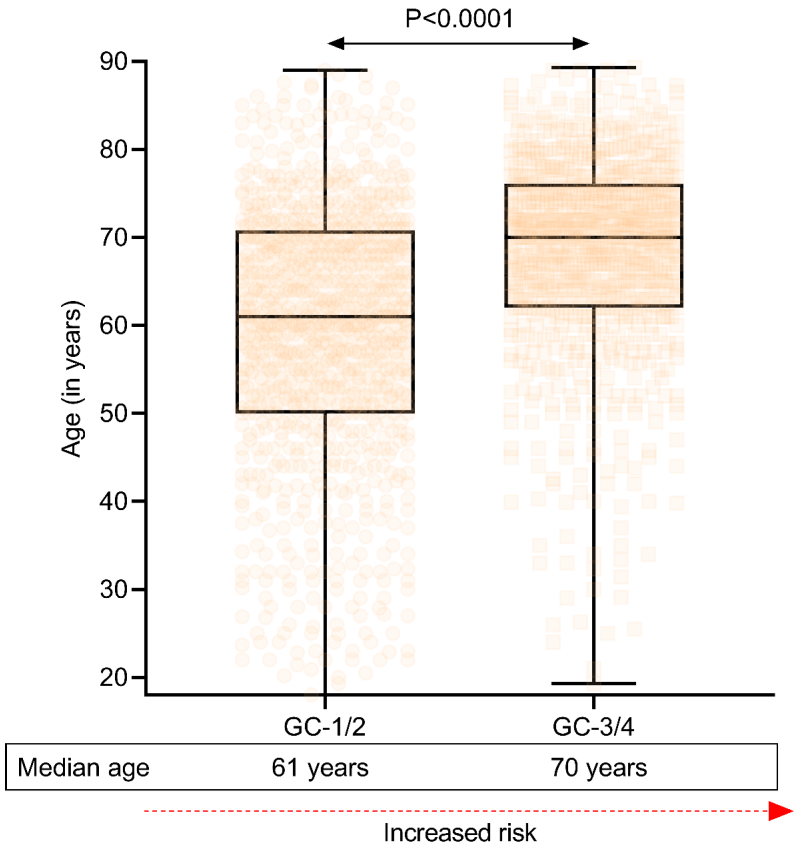


Figure A17: Age distribution per genomic clusters. The plot represents the comparison of age (in years) between Genomic cluster-1/2 (GC-1/2) vs. Genomic cluster-3/4 (GC-3/4). Levels of statistical significance is indicated using p-value.

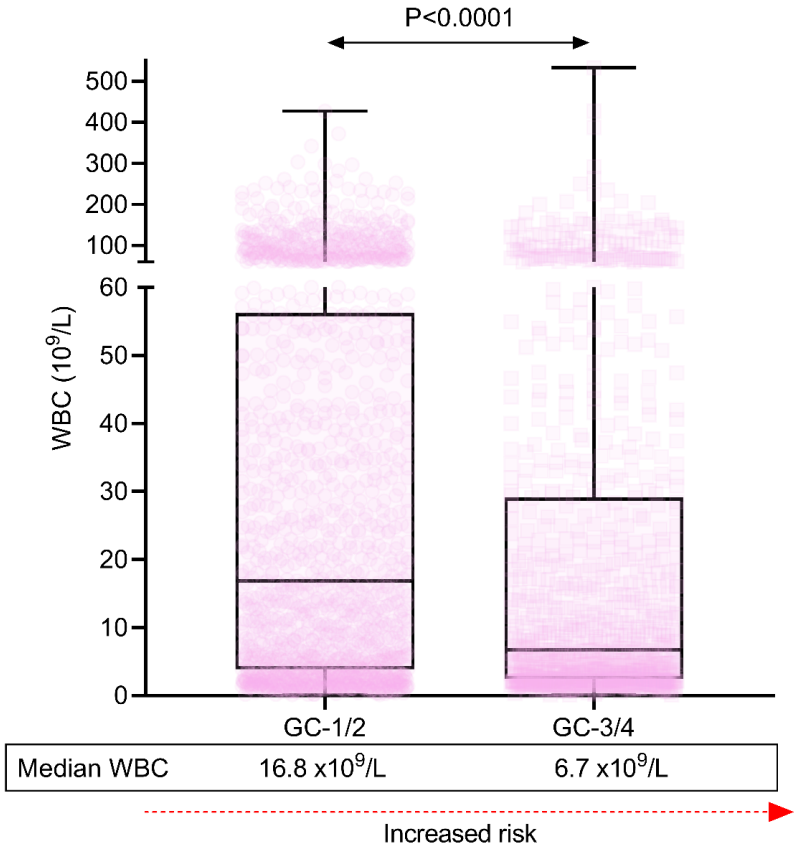


Figure A18: White blood cell count per genomic clusters. The plot represents the comparison of white blood cell count (WBC, in 109/L) between Genomic cluster-1/2 (GC-1/2) vs. Genomic cluster-3/4 (GC-3/4). Levels of statistical significance is indicated using p-value.

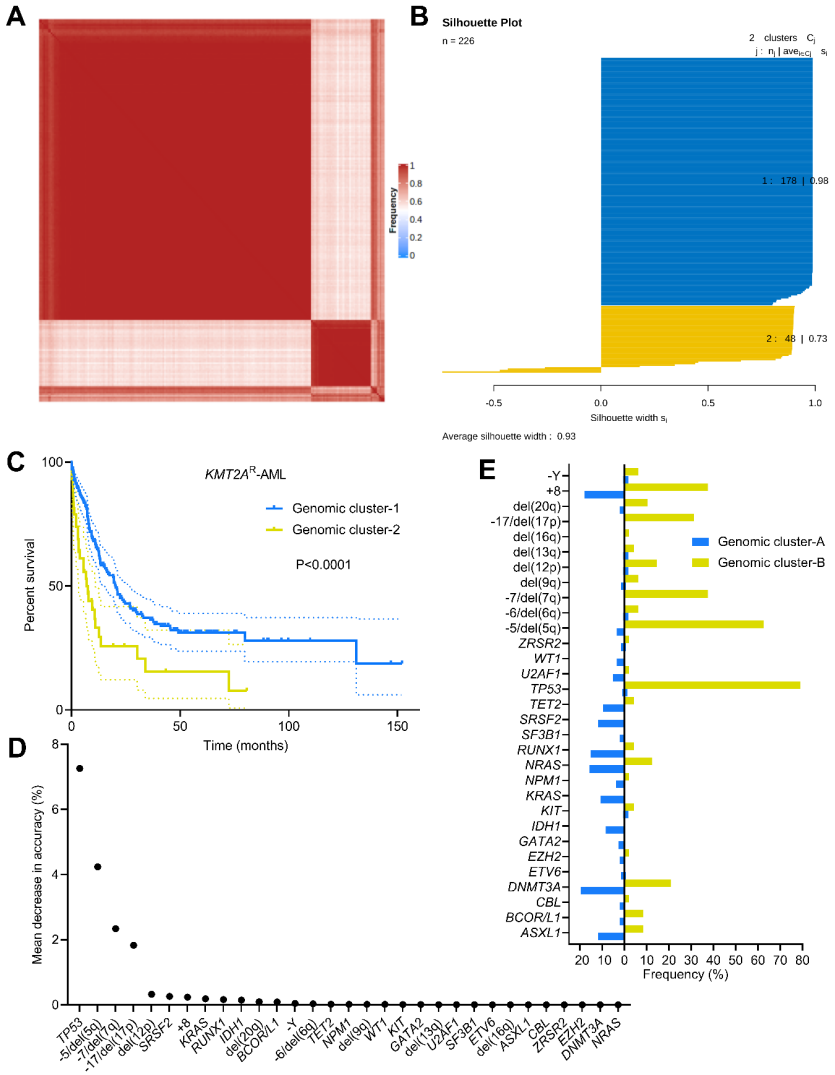


Figure A19: Novel genomic clusters of KMT2A-rearranged acute myeloid leukemia (KMT2A^R-AML) identified by unsupervised analyses. (A) Consensus matrix generated by applying latent class analysis on 1000 subsamples representing the frequency of two observations being clustered in the same group. (B) The plot represents the silhouette values in each of the identified clusters. Genomic cluster-1 (GC-1) in blue and genomic cluster-2 (GC-1) in yellow. (C) Kaplan-Meier analysis showing the overall survival (in months) of each cluster (1-2). (D) The bar plots representing the mutational profiles (described by the % frequency of genomic features) of GC-1 and GC-2 KMT2A^R-AML. (E) A plot showing the genomic features used in our model and their respective importance calculated by mean decrease in accuracy. The y-axis shows the decrease in overall classification accuracy if the given variable is removed from the model.

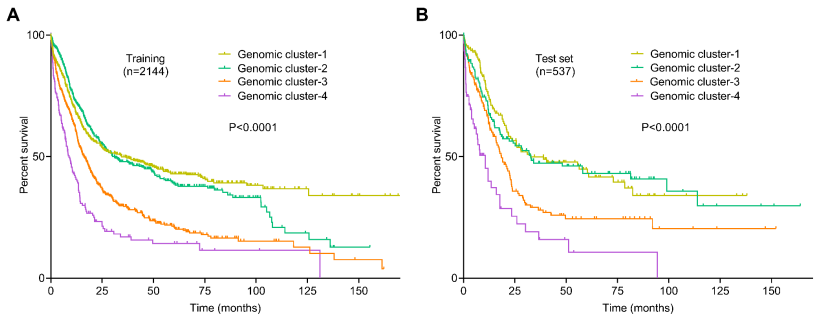


Figure A20: Internal validation: survival results in the training and test datasets. The training dataset contained 80% of the original cases (n=2144) that were randomly selected. Bayesian latent class analysis followed by random forest classification were applied to the training dataset. The test dataset contained 20% of the original cases (n=537) that were randomly selected. Random forest classification was applied to the test dataset. (A-B) Kaplan-Meier survival (using log-rank test) was used to plot survival curves of each genomic cluster in the training (A) and test (B) datasets. Levels of statistical significance are indicated using p-value.

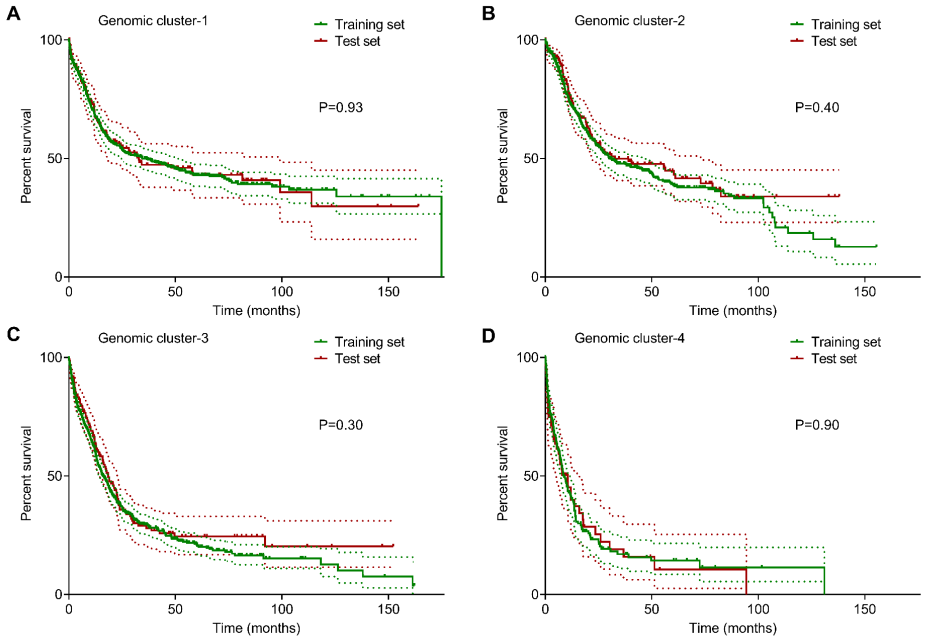


Figure A21: Results of the survival comparison of training and test datasets in the internal validation per each genomic cluster. Kaplan-Meier survival (using log-rank test) was used to plot and compare survival curves of the training and test sets per each genomic cluster (Panel A: genomic cluster-1, Panel B: genomic cluster-2, Panel C: genomic cluster-3 and Panel D: genomic cluster-4) as represented in the figure. Levels of statistical significance are indicated using p-values.

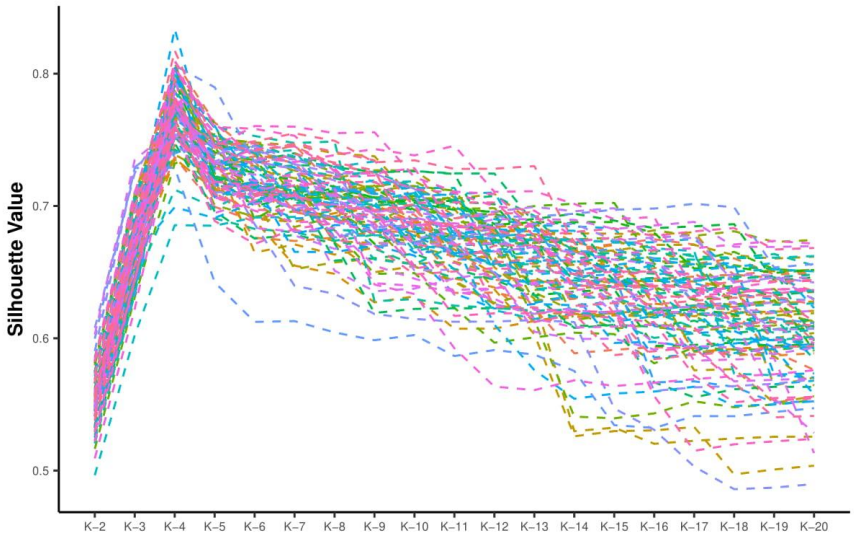


Figure A22: Validation of the selected number of genomic clusters. The plot represents the silhouette value with respect to the number of clusters that can be identified by Bayesian latent class analysis in 75% of our cohort. The plot shows that even when the number of patients was randomly reduced, BLCA did reproduce 4 clusters that attributed to the highest silhouette value. Therefore, the selection of 4 clusters based on the silhouette value can be further validated even in a smaller population of patients.

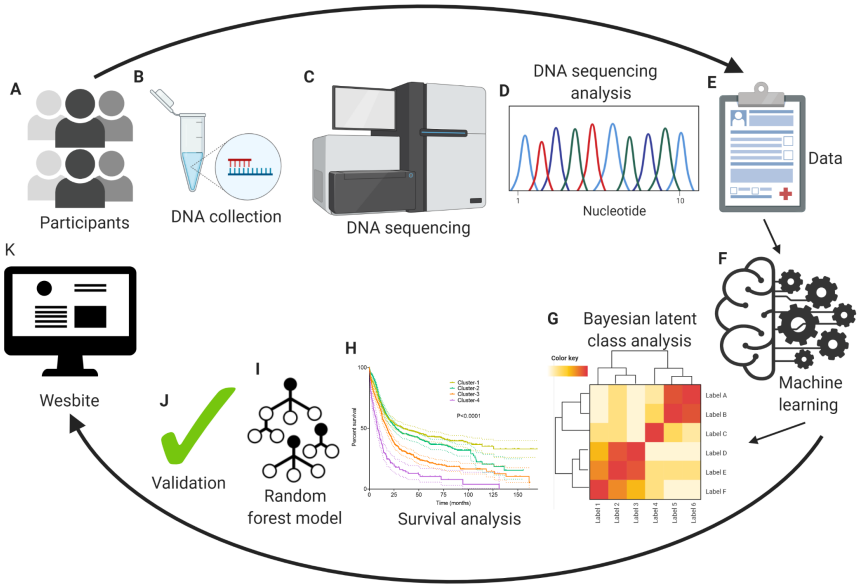


Figure A23: Conceptual figure. A schematic framework that illustrates our overall approach in this study.

A.0.3 Supplementary Methods

Genetic studies. For the data collected at CC, whole-exome sequencing (WES) was performed and paired disease and germline DNA of purified CD3+ lymphocytes were used. Whole-exome capture was accomplished by hybridizing sonicated genomic DNAs to a bait cDNA library synthesized on magnetic beads (SureSelect Human All Exon 50Mb or V4 kit, Agilent Technologies). Captured targets were sequenced using a HiSeq 2000 (Illumina) and standard protocols for 100-bp paired-end reads. Reads were aligned to the human genome (hg19) by a Burrows-Wheeler aligner (http://bio-bwa.sourceforge.net/) using a GATK pipeline that also extracted candidate variants/polymorphisms to reduce sequencing errors. Data were validated using targeted sequencing. Targeted sequencing was performed using a TruSeq Custom Amplicon kit (Illumina); a panel of 44 genes was interrogated (Table.S2). Sequencing libraries were generated according to an Illumina paired-end library protocol. The enriched targets were sequenced using a HiSeq 2000 or MiSeq (Illumina), at 862x coverage. Variants were annotated using Annovar14 and filtered by removing: i) synonymous single nucleotide variants; ii) variants only present in 140 unidirectional reads; and iii) variants in repetitive genomic regions. Variants with minimum depth less than 20 or number of high-quality reads less than 5 were filtered out. A bio-analytic pipeline developed in-house1-3 identified somatic mutations using sequences derived from controls and mutational databases such as dbSNP138, 1000 Genomes or ESP 6500 database, and Exome Aggregation Consortium (ExAC). Variant allelic frequencies (VAFs) were adjusted according to zygosity and copy number based on conventional metaphase karyotyping and/or single nucleotide polymorphism array results. Patients from the MLL cohort were investigated by NGS using different methods and gene panels as previously described.4-8 The gene sequencing methods of publicly-shared AML patients were previously described.9-11

Conventional cytogenetics. Metaphase cytogenetics was performed on BM aspirates. The median number of metaphases analyzed was 20. Chromosomal preparation was performed on G-banded metaphase cells using standard techniques, and karyotypes were described all

the patients according to the International System for Human Cytogenetic Nomenclature.12 **Clonal hierarchy.** The clonal hierarchy was resolved using our in-house designed VAF-based bioanalytic method and previously confirmed by the PyClone pipeline, which showed a high level of concordance.1-3 We assigned the clonal hierarchy by using VAFs (adjusted for copy number and zygosity) and then ranked the mutations. A mutation with the highest VAF that is at least 5% more than the 2nd highest VAF in each sample was defined as an "ancestral/dominant" mutation; those with less than 5% difference from the highest VAF were defined "ancestral/codominant" while those with VAFs of more than 5% difference from the highest VAF were considered "subclonal/secondary mutations". **Statistics.** Fisher's exact test and Chi-square test were used to compare categorical variables. Mann–Whitney U test/ Wilcoxon rank-sum test were used for continuous variables. All p- values were two-sided; those less than 0.05 were considered statistically significant. Univariate and multivariate Cox model analyses were also performed. All statistical computations were performed using R 3.6.2 (www.r-project.org) and Prism (GraphPad). In order to assess prognostic differences among the identified clusters, pairwise survival analysis using Kaplan-Meier estimator and log-rank test was performed. **Logistic regression.** Univariate and Multivariate logistic regression were applied in order to identify and compare prognostic genomic markers in pAML and sAML patients. Variables and patients with more than 80% of missing values were removed. Remaining missing values were imputed using R package missForest.13 **Survival analysis.** In order to assess prognostic differences among the identified clusters, we have performed pairwise survival analysis using Kaplan-Meier estimator and log-rank test. Bayesian Latent Class analysis for unsupervised clustering. We aimed to identify clinically 'functional' clusters for the AML patients given the binary data of mutation status for a panel of 44 previously determined genes and cytogenetic abnormalities by utilizing Bayesian Latent Class Analysis (BLCA), a finite mixture modeling framework using R package BayesLCA. The posterior distribution for the data, given the unobserved latent variables, can be written as:

$$p\left(X,Z\right)=\prod_{i=1}^{N}\prod_{g=1}^{G}\tau_{g}^{Z_{ig}+\delta_{g}-1}\prod_{m=1}^{M}\theta_{gm}^{X_{im}Z_{ig}+\alpha_{gm}-1}\left(1-\theta_{gm}\right)^{(1-X_{im})Z_{ig}+\beta_{gm}-1}$$

Where τ is the probability of belonging to a class g, θ is a GxM dimensional matrix being the probability of a feature $i = 1$ given the class membership. Given a predefined number of latent variables, we estimated the parameters using EM framework. Since the number of latent classes are predefined, we applied the model with different number of latent classes and selected the best model using Bayesian Information Criterion (BIC). Estimated parameters were then used to calculate the posterior probability of samples belonging to clusters $P\left(z_i = k \vee x_i, \theta\right)$, hence the samples were partitioned into different clusters by selecting the highest probability cluster. In order to account for outlier observations, we applied this approach 1000 times to different subsamples of observations (%75 sampling) to generate 1000 different clustering schemes. Thousand clustering schemes were then aggregated into a consensus matrix by calculating the frequency of assignment into the same cluster for all pairwise comparisons across 1000 iterations. Hyperparameter tuning for interaction depth was done using 10-fold cross-validation where we set the parameter to 3.

Normalization of primary and secondary acute myeloid leukemia distributions. In order to assess the distribution of pAML and sAML among the identified clusters with an unbiased approach, we normalized pAML and sAML distributions with randomly sampling equal number of pAML and sAML patients for 10000 times and evaluated the probability of observing pAML in each cluster (**Figure 2D**).

Random Forest for extraction of genomic features. Secondary aim of unsupervised clustering was to extract the relevant genomic features that facilitated molecular based classification of patients and building a genomic classification model. For this purpose, we proposed to use off-the-shelf machine learning method random forests as a multiclass classification problem to classify the clusters initially identified via BLCA framework.

Random forest is a model averaging technique with base models chosen as trees applied successfully to wide variety of problems including cancer. The method combines output from pre-selected number of trees applied to a subset of original data with both features and observations being randomly selected. This procedure eliminates the pairwise correlation of trees. To assess the importance of individual variables, mean decrease in accuracy was used. However, since random forests employ sampling strategy, each model can produce slightly different importance measure for each variable, hence we performed the procedure 100 times and plot the distributions. We also generated cluster-wise importance measures by removing each variable from the model and calculating the mean decrease in accuracy for the specific cluster (**Figure 3A-E**). The hyperparameter selection of the depth of trees is done using 10-fold cross-validation with the total number of trees set to 1500.

Uniform resource locator. (URL: `https://drmz.shinyapps.io/local_app/`)

Appendix B

MOLECULAR PATTERNS IDENTIFY DISTINCT SUBCLASSES OF MYELOID NEOPLASIA

B.0.1 Supplementary Tables

Table B1: **Summary of the sources of myelodysplastic syndrome and secondary acute myeloid leukemia cases included in our study**

Cohorts	Total number of patients
Our cohorts	2902
Cleveland Clinic Foundation (CCF)	1627
Munich Leukemia Laboratory (MLL)	1275
Public cohorts	686
Beat AML master trial	45
Euro-MDS cohort2	641
External Validation Cohorts	419
Wayne State University Karmanos Comprehensive cancer center	207
University of Texas Southwestern Simmons Comprehensive Cancer Center	212

Table B2: List of 40 genes in our targeted panel used for the molecular machine learning model

ASXL1	BCOR	BCORL1	CALR	CUX1	CEBPA	CBL	CSF3R
DNMT3A	DDX41	EZH2	ETV6	FLT3	GATA2	GNAS	IDH1
IDH2	JAK2	KRAS	KIT	MPL	NRAS	NPM1	NF1
NOTCH1	PTPN11	PHF6	RAD21	RUNX1	SF3B1	SRSF2	SMC1A
SMC3	STAG2	SETBP1	TET2	TP53	U2AF1	WT1	ZRSR2

Table B3: **Clinical, cytogenetic, and molecular characteristics of original and validation cohorts**

Variables	Experimental Cohort n=3588	External Cohort n=.412	P-value
Age, median (IQR)	72 (64-77)	69 (62-75)	0.00
Gender			0.81
Male, n (%)	2143 (60)	248 (60)	
Female, n (%)	1444 (40)	163 (40)	
BM blast %, median (IQR)	4 (2-13)	7 (2-19)	0.00
Diagnosis			0.00
LR-MDS	2079 (58)	156 (38)	
HR-MDS	774 (22)	134 (33)	
s-AML	735 (20)	122 (30)	
Cytogenetics			0.00
Normal	2023 (57)	252 (62)	
Abnormal	1548 (43)	156 (38)	
Number of MT			0.00
0	825 (23)	33 (8)	
1-2	1666 (46)	204 (49)	
3-4	813 (23)	114 (28)	
>4	581 (16)	61 (15)	
Molecular clusters			
1	201 (6)	25 (6)	
2	920 (26)	74 (18)	
3	76 (2)	2 (1)	
4	313 (9)	17 (4)	
5	107 (3)	10 (2)	
6	301 (8)	19 (5)	
7	225 (6)	54 (13)	
8	236 (7)	13 (4)	
9	219 (6)	26 (6)	
10	143 (4)	9 (2)	
11	121 (3)	16 (4)	
12	130 (4)	19 (5)	
13	391 (11)	118 (29)	
14	205 (6)	10 (2)	
Status Death, n (%)	1559 (44)	190 (46)	0.32

Table B4: **Clinical, cytogenetic, and molecular characteristics of all risk groups**

Variables	All	Low.Risk	Low-Int.Risk	High-Int.Risk	High.Risk	Very.High.Risk
Total population	3588	456	1457	962	322	391
Test cohort	718 (20)	89 (19)	291 (20)	192 (20)	59 (18)	87 (22)
Training cohort	2870 (80)	367 (81)	1166 (80)	770 (80)	263 (82)	304 (78)
Age, median (IQR)	72 (64-77)	73 (68-78)	71 (63-77)	72 (64-78)	73 (65-78)	71 (63-77)
Gender						
Male	2143 (60)	258 (57)	793 (54)	674 (70)	201 (62)	217 (56)
Female	1444 (40)	198 (43)	664 (46)	287 (30)	121 (38)	174 (45)
Labs						
WBC (109/L)	4 (3-11)	6 (4-8)	5 (3-10)	6 (3-15)	5 (3-15)	3 (2-8)
Hb(g/dL)	10 (9-11)	10 (9-11)	10 (9-12)	10 (9-11)	9 (8-11)	10 (9-11)
Platelets (109/L)	112 (50-240)	290 (169-395)	112 (60-227)	102 (44-202)	67 (27-158)	58 (30-102)
BM blast %	4 (2-13)	2 (1-4)	4 (2-11)	5 (2-14)	11 (4-32)	12 (3-22)
Diagnosis						
LR-MDS	2079 (58)	380 (83)	902 (62)	528 (55)	119 (37)	150 (38)
HR-MDS	774 (22)	53 (12)	285 (20)	242 (25)	83 (26)	111 (28)
sAML	735 (21)	23 (5)	270 (19)	192 (20)	120 (37)	130 (33)
Cytogenetics						
Normal	2023 (57)	455 (100)	1220 (84)	347 (37)	1 (0)	0 (0)
Abnormal	1548 (43)	1 (0)	237 (16)	600 (63)	319 (100)	391 (100)
Number of MT						
0	825 (23)	0 (0)	521 (36)	138 (14)	74 (23)	92 (24)
1-2	1666 (46)	298 (66)	658 (45)	377 (39)	114 (35)	219 (56)
3-4	813 (23)	138 (31)	220 (16)	302 (31)	89 (28)	64 (17)
>4	284 (8)	20 (4)	58 (4)	145 (15)	45 (14)	16 (4)

B.0.2 Supplementary Figures

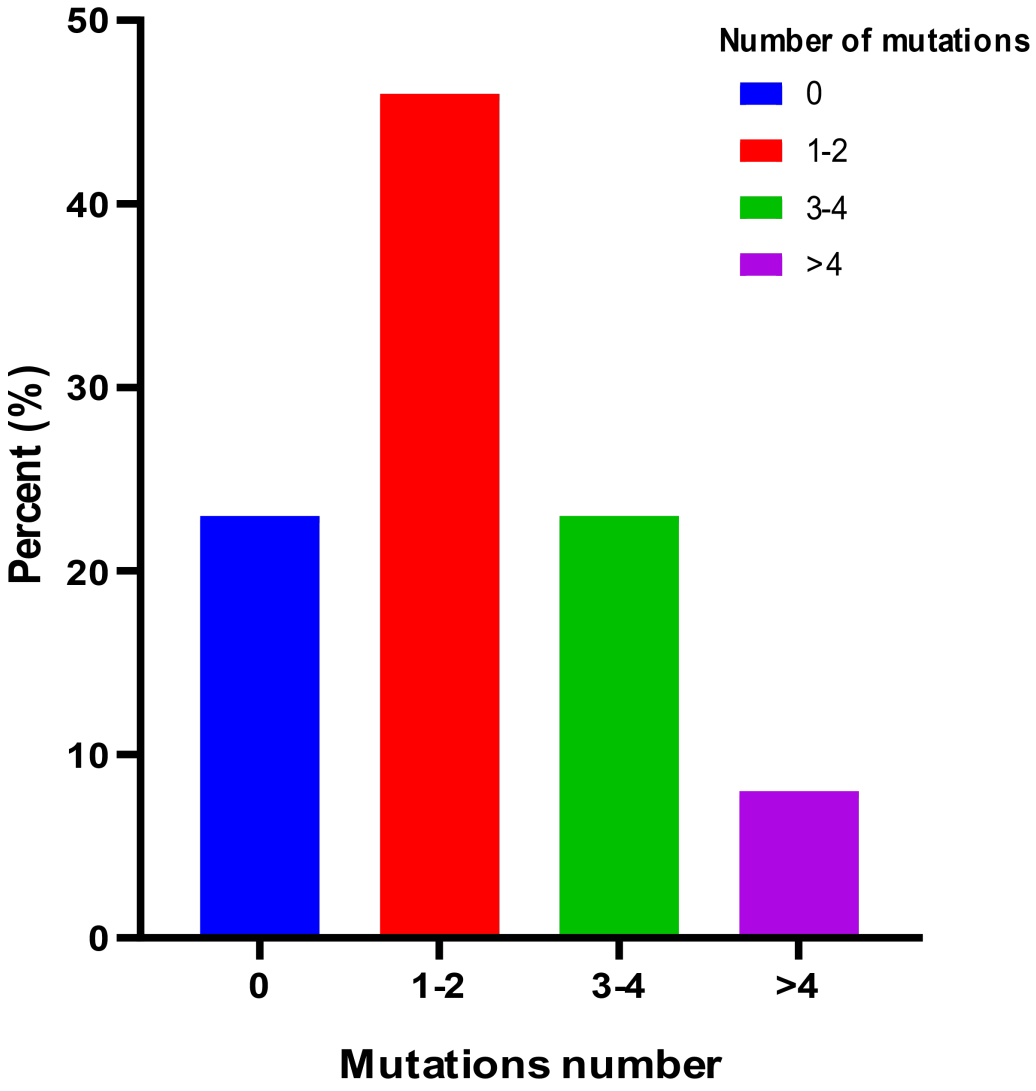


Figure B1: Frequency of total mutations number as distributed among our myelodysplastic syndrome (MDS) and secondary acute myeloid leukemia (sAML) cases.

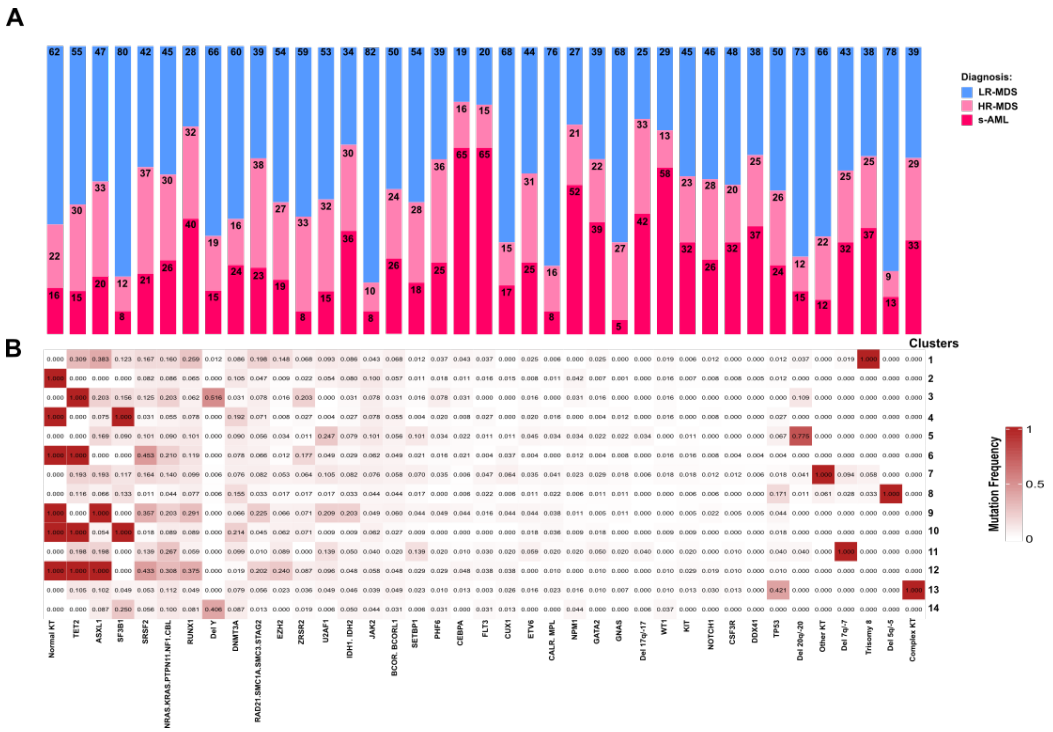


Figure B2: **(A)** Histogram bars represent the distribution of molecular hits and cytogenetics abnormalities among LR-MDS, HR-MDS, and sAML patients illustrated by a specific figure color legend. **(B)** Heatmap representation of the frequency of molecular mutations and cytogenetic abnormalities per each genomic cluster.

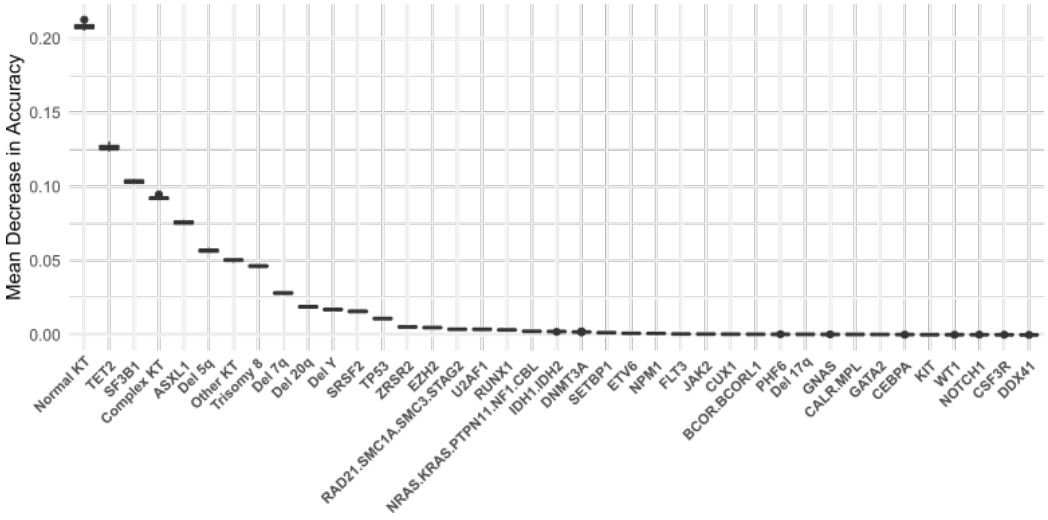


Figure B3: Genetic features ordered by 'global importance' measured by mean decrease in accuracy for the random forest classification model. A mean decrease in accuracy ≥ 0.01 was considered significant

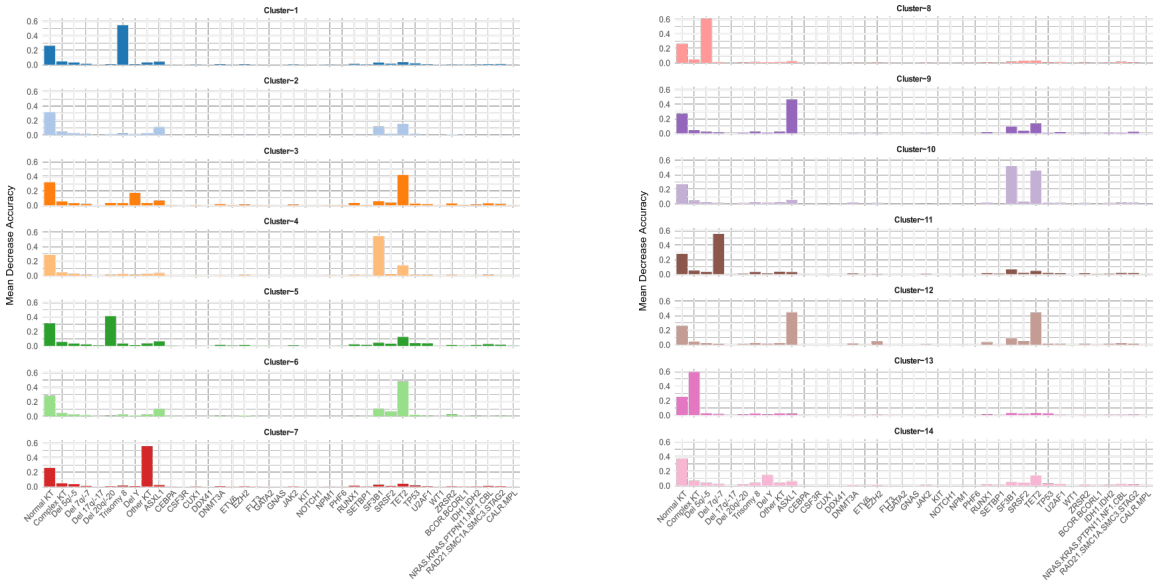


Figure B4: Cluster-specific importance of genetic features measured by mean decrease in accuracy for the random forest classification model. A mean decrease in accuracy0.01 was considered significant.

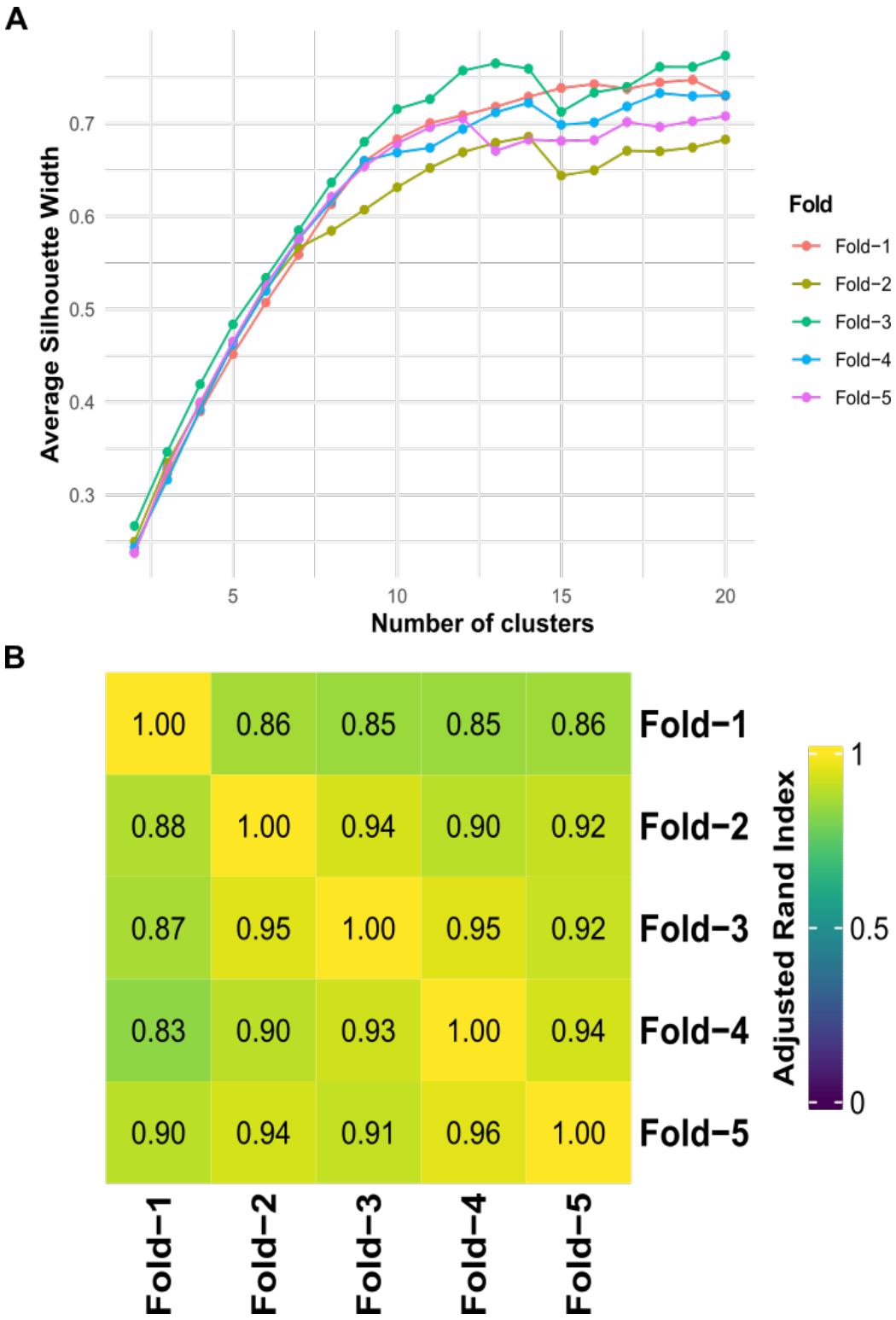


Figure B5: K-fold cross-validation method for the proposed unsupervised clustering approach. A: The figure represents the silhouette values based on the number of the clusters. Total cluster number of 14 was associated with highest silhouette values in all folds. B: Overlap between the sub-groups (folds) based on the predicted assignments of random-forest classification models generated from each fold separately. More specifically, row j comparing column k shows the overlap of cases in fold j using Adjusted Rand Index (ARI) classified by the model trained on fold j only.

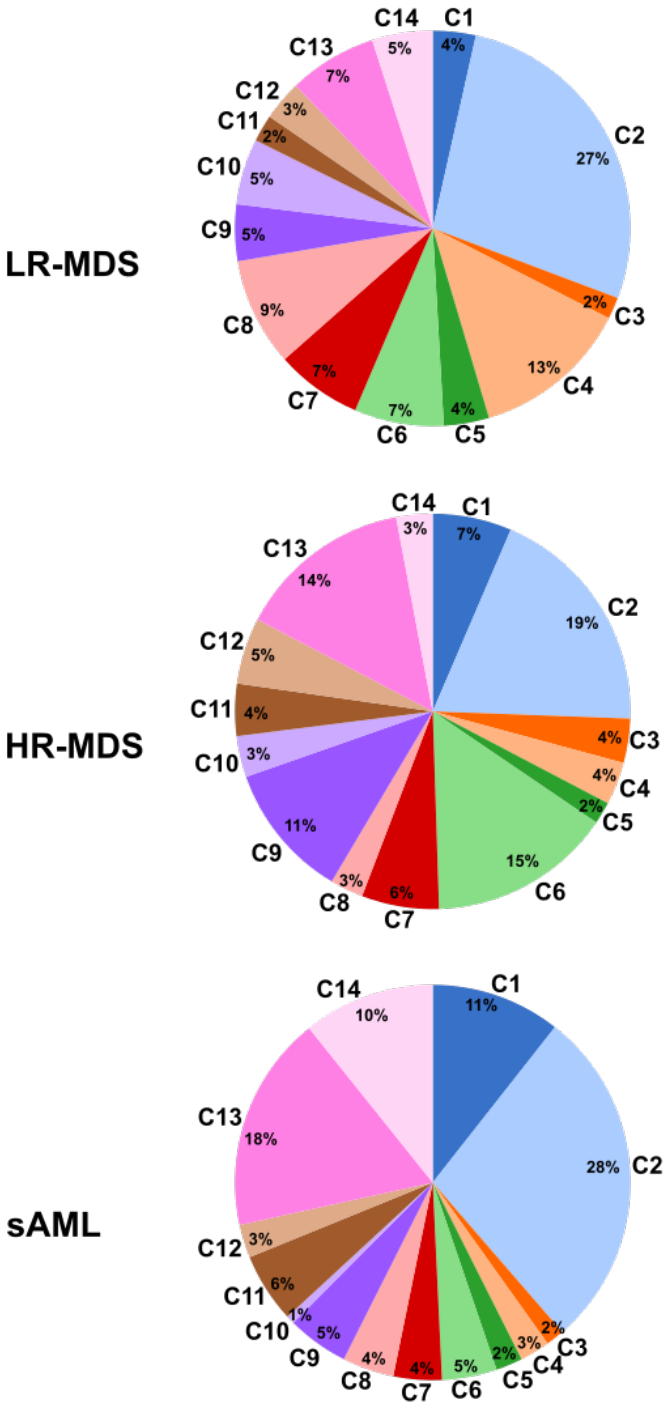


Figure B6: Molecular clusters (C) percentage in low-risk myelodysplastic syndrome (LR-MDS), high-risk myelodysplastic syndrome (HR-MDS), and secondary acute myeloid leukemia (sAML) patients. The pie charts demonstrate the percentage of each molecular clusters in different clinical diseases. Each molecular cluster is presented by a specific figure legend color.

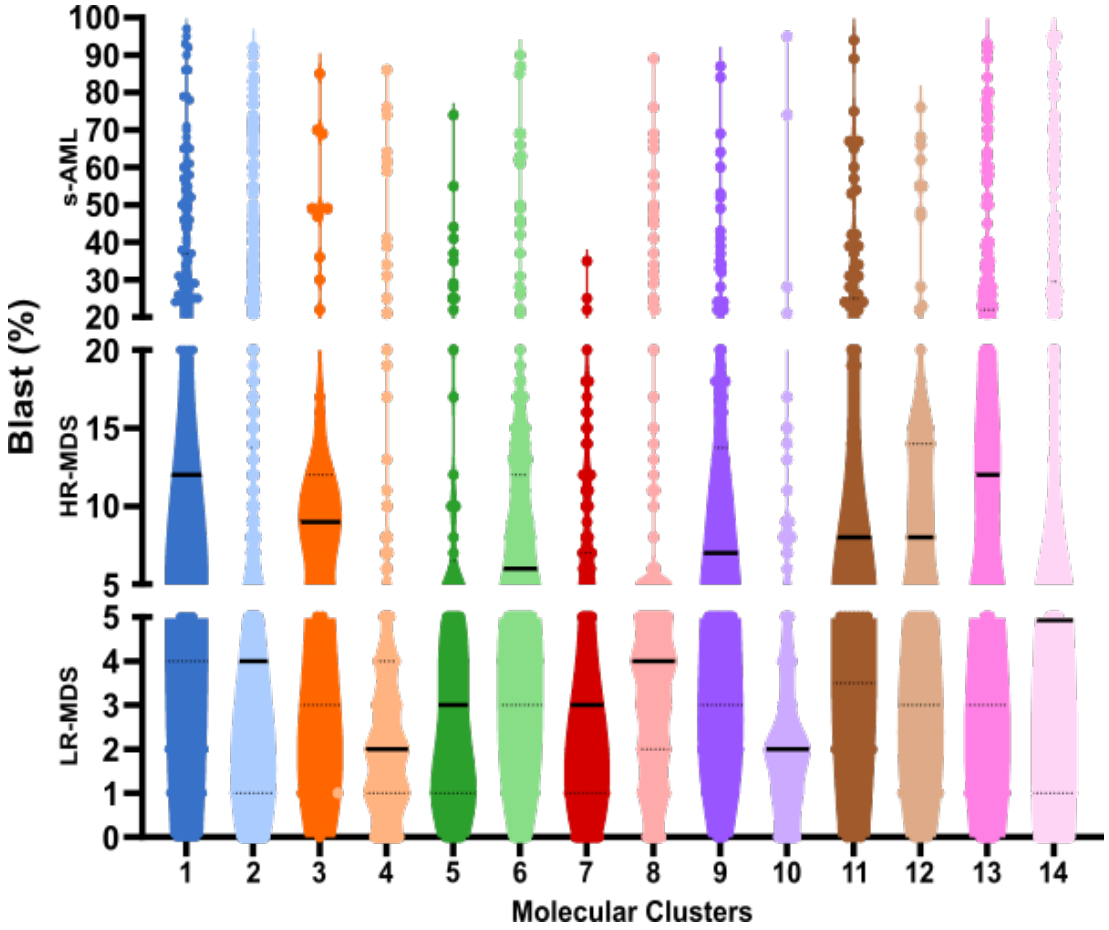


Figure B7: Bone marrow blast percent (%) per molecular clusters. The plot represents the distribution of bone marrow blast percent in each molecular cluster. Solid lines represent median and dashed lines represent the 95% confidence intervals.

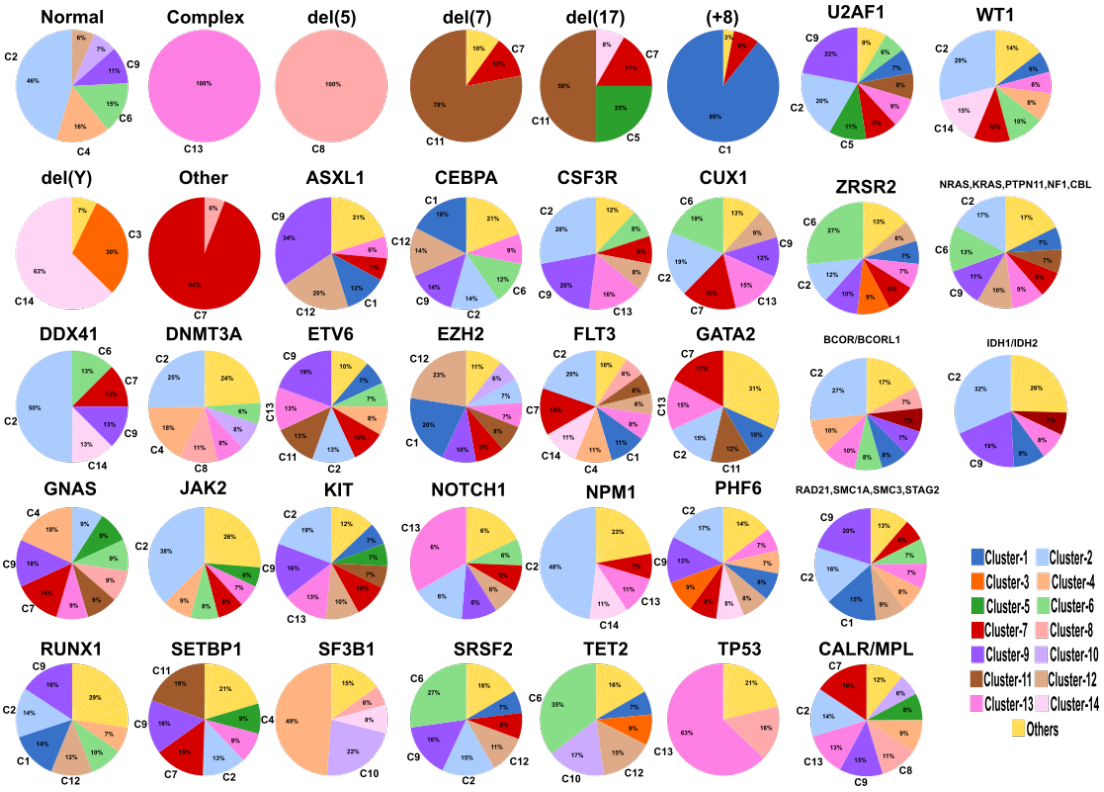


Figure B8: Distribution of all the molecular mutations and cytogenetic abnormalities used to build our scheme across molecular Clusters (C). The pie charts illustrate the abundance of each molecular cluster (C) with regards to gene mutations and cytogenetic abnormalities. Each molecular cluster is presented by a specific figure legend color.

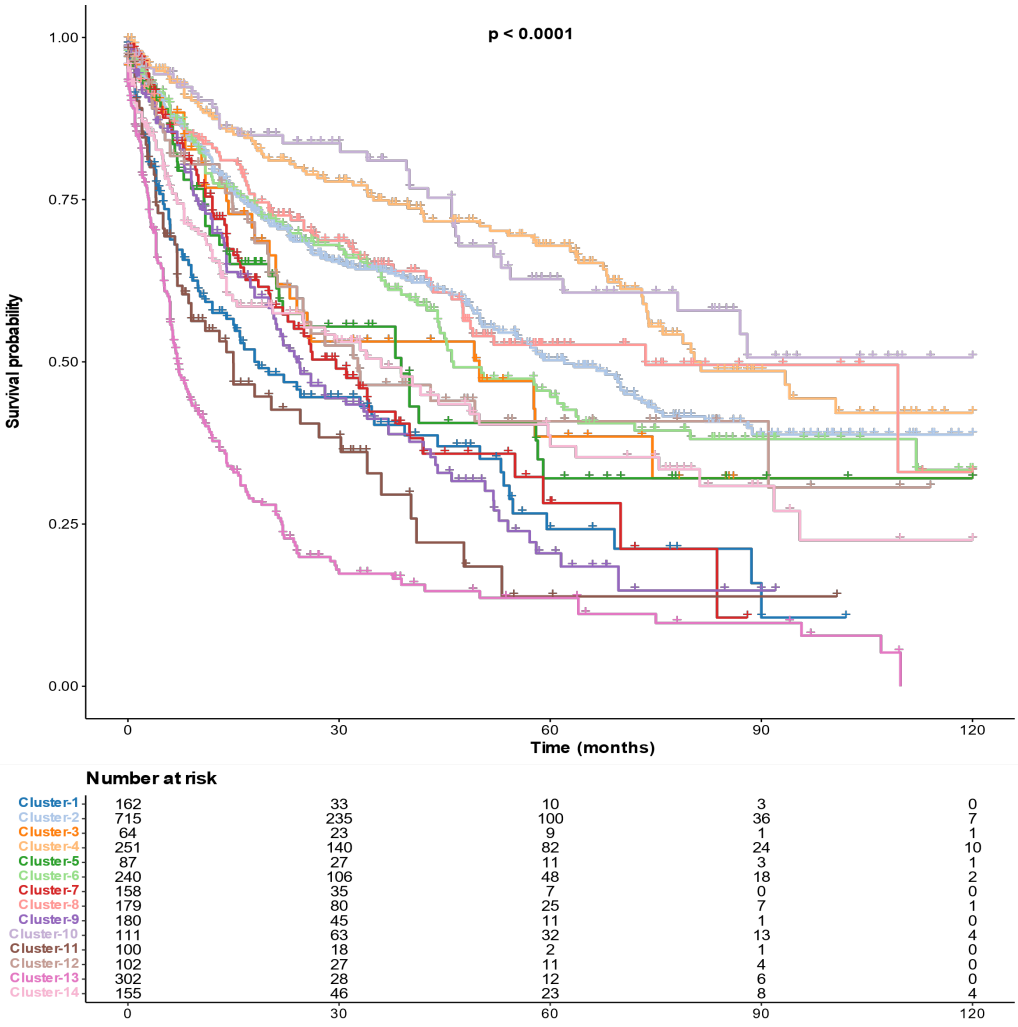


Figure B9: Kaplan-Meier analysis showing the overall survival (in months) of cases assigned to different molecular clusters (cluster-1 to cluster-14). Statistically significant difference of log-Rank test is indicated by the p-value.

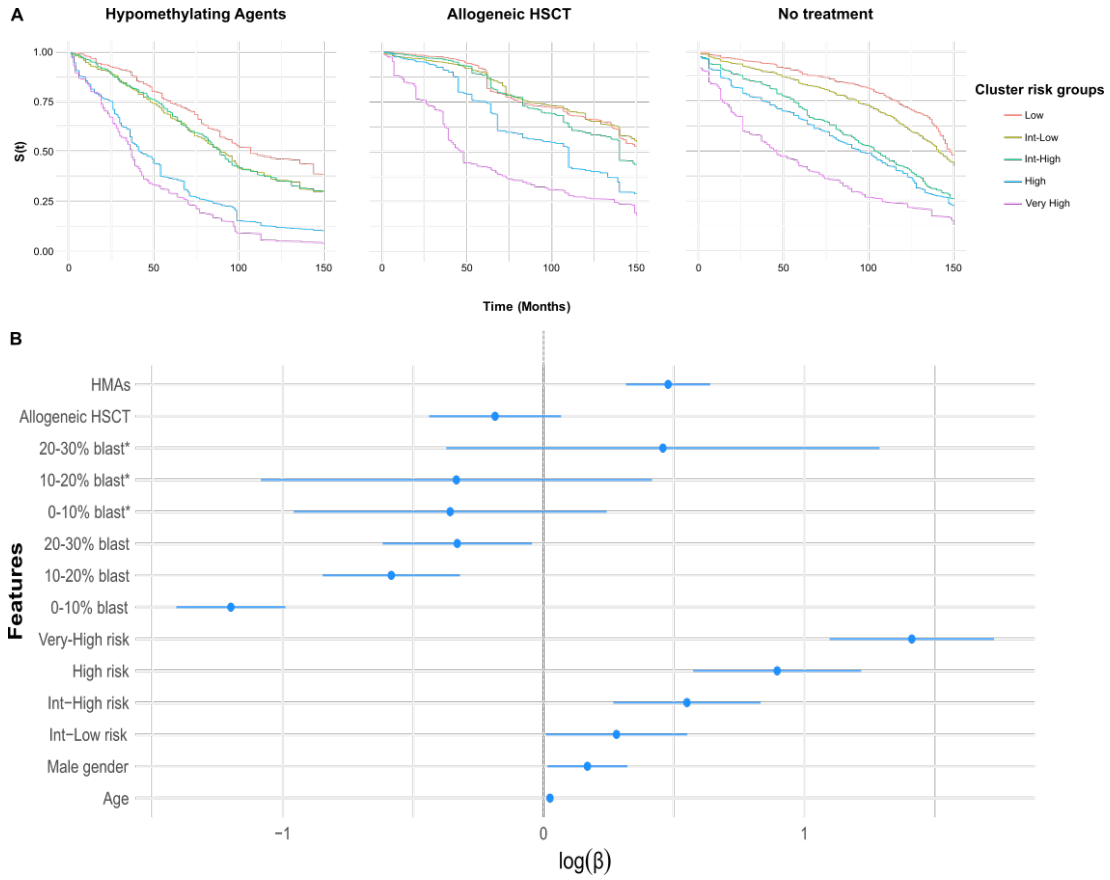


Figure B10: **(A)** Non-parametric survival estimation using Random Survival-Forest for different genomic risk groups adjusted for hypomethylating agents (HMAs) treatment, allogeneic hematopoietic stem cell transplant (HSCT), no treatment, age and sex. Survival curves are estimated for a pseudo-patient (male, aged 75 years) showing the effect of molecular clusters adjusting for treatment and other clinical variables. Each risk group is presented by a specific figure legend color. **(B)** Subgroup analysis of overall survival according to age, gender, cluster risk groups, bone marrow blast percent before 25 months and after 25 months (asterisk [*]), HMAs treatment, and allogeneic HSCT.

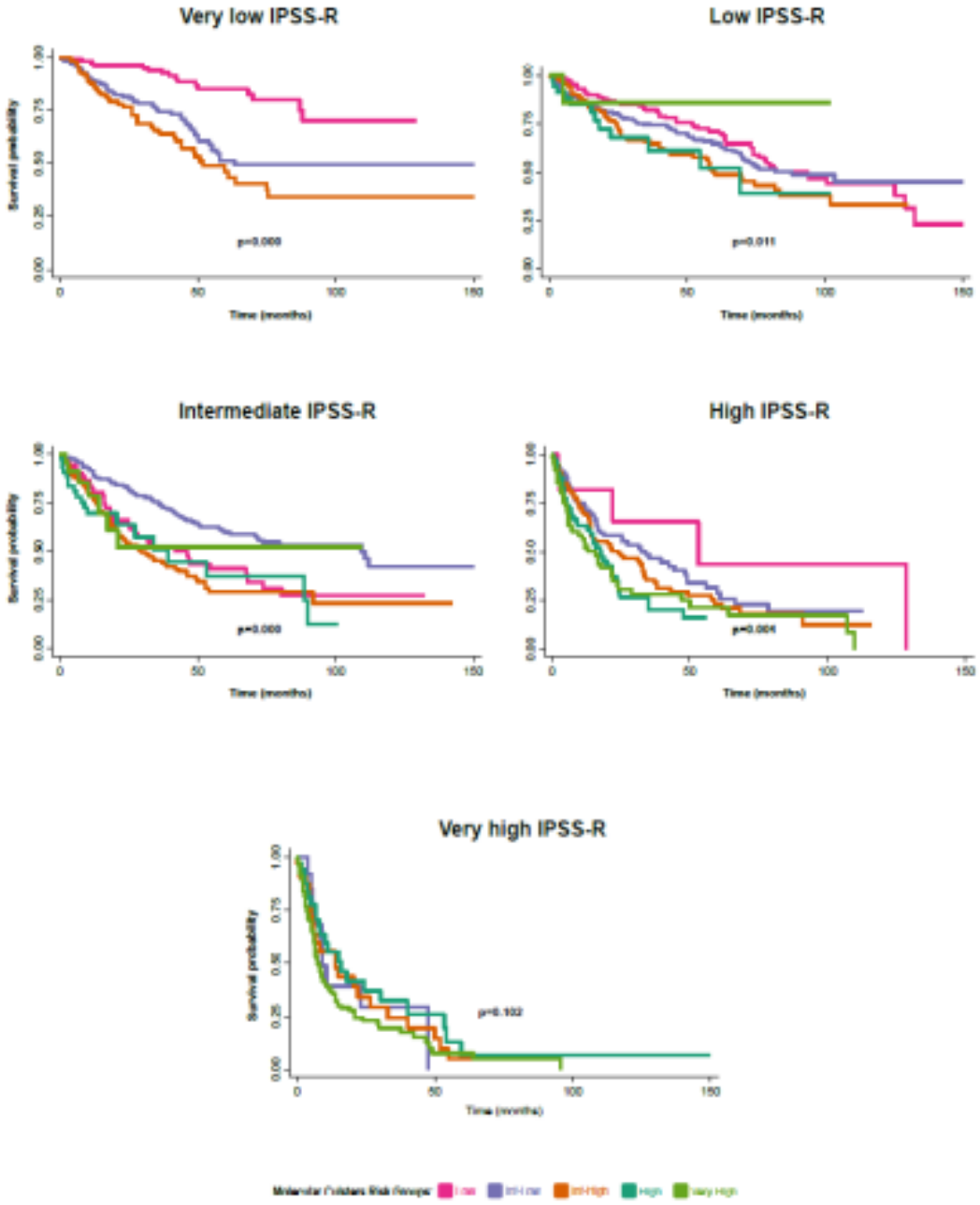


Figure B11: Kaplan-Meier analysis showing the overall survival (in months) of cases assigned to different molecular risk groups (Low, Int-low, Int-high, High, and Very-high) among different Revised International Prognostic Scoring System (IPSS-R) risk groups. Statistically significant difference of log-Rank test is indicated by the p-value.

B.0.3 Supplementary Results

B.0.3.1 Examples of molecular associations in our MCs

1. *SF3B1* mutations were enriched in MC4, MC10, and MC14 and associated with *TET2* (100%) mutations in MC10, indicating functional relationships. Notably, this group was distinct from MC14 in which *SF3B1* mutations were grouped with delY and *SRSF2* mutations. The diverse pathobiological derivation of these groups was also illustrated by their significant survival differences. Other splicing factor mutations were instead functionally distinctive and were assigned to different genomic groups. Indeed, *SRSF2* mutations were enriched in MC9 and MC12 and associated with ASXL1 mutations and normal karyotype, but MC12 cases had additional *TET2* mutations. A similar principle seemed to apply to *U2AF1* mutations, abundant in MC5. These results argue that the genomic sub-classification of splicing factor mutations is strongly dependent on the presence/absence of other unique correlating cofounders affecting the functionality and biology of the distinct cluster.

2. Traditionally, epigenetic modulator mutations have been grouped together for classification purposes, often ignoring their distinct or even often opposite function[8, 85, 155]. In contrast, our ML-derived model highlights these functional differences, underscoring intertwining relationships across different molecular pathways of leukemogenesis. For instance, we found that *IDH1/IDH2* mutations were mainly abundant in two MCs (MC2 and MC9), which had discrete survival differences and a significantly higher percentage of *STAG2* and *IDH1/IDH2* mutations in MC9 vs. MC2, in which they coincided with more frequent *RAS* and *DNMT3A* hits. Similarly, the functional effect of *DNMT3A/SF3B1* co-mutations was reflected by a better overall-survival of MC4 and MC10 (Low-Risk group). This phenomenon can be functionally explained by the relative mitigation effects of *SF3B1* mutation on *DNMT3A* clones[156, 157]. Because *DNMT3A* and *TET2* have different biological and possibly opposite func-

tions, mutations in these epigenetic regulators belong to distinct functional pathways and separate MCs. Both *EZH2* and *UTX* possess opposite H3K27 methylation effects and thus their mutations clustered separately and associated/substituted with/for distinct cluster-defining hits. Finally, *RUNX1* mutations were abundant in MC9, MC12 and MC1 and were significantly associated with either *ASXL1* and *SRSF2* mutations as previously reported[69].

3. While our results reaffirm previous studies regarding the poor survival outcome associated with complex cytogenetics and TP53 mutations[8, 158–160], the impact of *TP53* allelic configuration on pathology and prognosis was also reflected in our molecular clustering, e.g., *TP53* mutations in MC8 were mostly monoallelic (70%), explaining the better outcomes as compared to other MCs with *TP53* in biallelic configuration. In addition, mutually exclusive *PPM1D* mutations coincided with *TP53* clusters (MC13 and MC8), pointing towards to the known similar pathogenic pathway[161–163].

B.0.4 Supplementary Methods

B.0.4.1 Genetic studies

For the data collected at CC, whole-exome sequencing (WES) was performed and paired disease and germline DNA of purified CD3+ lymphocytes were used. Whole-exome capture was accomplished by hybridizing sonicated genomic DNAs to a bait cDNA library synthesized on magnetic beads (SureSelect Human All Exon 50Mb or V4 kit, Agilent Technologies). Captured targets were sequenced using a HiSeq 2000 (Illumina) and standard protocols for 100-bp paired-end reads. Reads were aligned to the human genome (hg19) by a Burrows-Wheeler aligner (http://bio-bwa.sourceforge.net/) using a GATK pipeline that also extracted candidate variants/polymorphisms to reduce sequencing errors. Data were validated using targeted sequencing. Targeted sequencing was performed using a TruSeq Custom Amplicon kit (Illumina); a panel of 40 genes was interrogated (**Table** B2). Sequencing libraries were generated according to an Illumina paired-end library protocol. The enriched targets were sequenced using a HiSeq 2000 or MiSeq (Illumina), at 862x coverage. Variants were annotated using Annovar14 and filtered by removing: i) synonymous single nucleotide variants; ii) variants only present in 140 unidirectional reads; and iii) variants in repetitive genomic regions. Variants with minimum depth less than 20 or number of high-quality reads less than 5 were filtered out. A bio-analytic pipeline developed in-house[59, 60, 77] deidentified somatic mutations using sequences derived from controls and mutational databases such as dbSNP138, 1000 Genomes or ESP 6500 database, and Exome Aggregation Consortium (ExAC). Variant allelic frequencies (VAFs) were adjusted according to zygosity and copy number based on conventional metaphase karyotyping and/or single nucleotide polymorphism array results. Patients from the MLL cohort were investigated by NGS using different methods and gene panels as previously described[59, 60, 62]. The gene sequencing methods of publicly-shared patients were previously described[8, 56, 80, 81].

B.0.4.2 Conventional cytogenetics

Metaphase cytogenetics was performed on bone marrow (BM) aspirates. The median number of metaphases analyzed was 20. Chromosomal preparation was performed on G-banded metaphase cells using standard techniques, and karyotypes were described all the patients according to the International System for Human Cytogenetic Nomenclature[164, 165].

B.0.4.3 Statistical Methods

Fisher's exact test and Chi-square test were used to compare categorical variables. Mann–Whitney U test/ Wilcoxon rank-sum test were used for continuous variables. All p- values were two-sided; those less than 0.05 were considered statistically significant. All statistical computations were performed using R 3.6.2 (www.r-project.org) and Prism (GraphPad). To assess prognostic differences among the identified clusters, pairwise survival analysis using Kaplan Meier estimator and log-rank test was performed.

B.0.4.4 Autoencoder

Autoencoders are neural-network architectures, which can be designed to generate efficient compressed/low dimensional representations of given data. Here, we used a single layer autoencoder with shared layer between encoder and decoder to generate low-dimensional representations of binary mutation profiles of MDS samples. With a single shared layer, autoencoders can learn to capture the principal component space without orthogonality constraint, hence a proxy for binary-PCA analysis. We used TensorFlow framework to optimize via Adam optimizer and learning rate and batch size set to $1E-4$ and 32 respectively. l1 and l2 regularization parameters are set to 0.1 as well to prevent possible overfitting[166, 167].

B.0.4.5 Gaussian Mixture Model

Gaussian-mixture models (GMM) are model-based clustering methods where observations are separated into components/clusters representing gaussian distributions parameterized by different μ and σ. We used scikit-learn package to fit GMM with expectation-maximization over increasing number of components and used Bayesian Information Criterion (BIC) to select the number of clusters. Over 100 iterations with random sub-sampling of binary observation vectors of mutation profiles, we first generated a low dimensional embedding of the sub-sampled data and used GMM to cluster the observations, where the number of GMM components was selected using Bayesian Information Criterion (BIC). Keeping track of co-clustering of observations at each iteration, we generated a consensus-matrix representing the frequency of clustering observations in the same cluster. The generated consensus-matrix was further clustered using hierarchical-clustering with Ward's criteria to create the final cluster assignments[168].

B.0.4.6 Unsupervised Clustering

Coupling Autoencoders and Gaussian-Mixture Models, we generated unsupervised clusters of MDS cases via Consensus approach. Over 100 iterations with random sub-sampling of data, we embedded the observations using the single layer autoencoder and clustered using the gaussian-mixture model. Keeping track of co-clustering of observations we generated a consensus-matrix. Finally, the consensus-matrix is clustered using hierarchical clustering with Ward's criteria and Silhouette value to select the number of clusters to generate the final cluster assignments for all cases.

B.0.4.7 Validation

The model was internally and externally validated. Internal validation was performed by dividing the whole cohort into training (80%) and test (20%) sets. The model was developed based on the training set only. Survival analysis by random forest (RF) was then performed

on the remaining 20% test set separately for internal validation. External validation was conducted on an independent cohort of MDS and sAML patients from the University of Texas Southwestern and the Wayne State University (**Table** B1).

Appendix C

STABILITY OF SCRNA-SEQ ANALYSIS WORKFLOWS IS SUSCEPTIBLE TO PREPROCESSING AND IS MITIGATED BY REGULARIZED OR SUPERVISED APPROACHES

C.0.1 Supplementary Methods

C.0.1.1 Imputation

ScImpute: ScImpute utilizes mixture modeling strategy to determine genes and cells that require imputation simultaneously. Using a Gamma-Normal mixture model, ScImpute first determines the dropout probability for each gene in each cell subpopulation identified by prior clustering. Imputation is then conducted using non-negative least squares regression using expression values from similar cells. We have used default parameters with dropout threshold set to 0.5 and initial clustering is done using *quickCluster* function from *scran* package in R with clustering method set to *igraph* and minimum size set to %10 number of cells.

DrImpute: DrImpute generates imputed counts using clustering and expression averaging where the cluster configurations are bootstrapped to result in robust estimates. All the parameters are set to default values.

C.0.1.2 Normalization

ScTransform: ScTransform utilizes a modified negative binomial (NB) regression where regularization using kernel-smoothing across mean expression levels for NB parameters is used. Passing library size as a covariate to NB regression allows for efficient normalization while accounting for mean-variance relationship. We have used *sctransform* package with parameter number of genes set to use all the genes.

Deconvolution: Normalization with deconvolution strategy aims to utilize count informa-

tion from cells with similar transcriptional profiles. For this purpose, initially clustered cells are normalized against the population as a single pool of cells. The normalization factor is then deconvolved to generate cell-wise normalization factors using least squares methods. Similar to ScImpute, initial pool of cells are generated by using *quickCluster* method from *scran* package with clustering method set to *igraph* and minimum size set to %10

Deep Count Autoencoder: Deep Count Autoencoder (DCA) follows a slightly different strategy in which autoencoder based neural network is used to estimate dropout and dispersion parameters with likelihood based on negative-binomial (NB) or zero-inflated negative-binomial (ZINB) models are used. DCA aims to identify the latent structure in scRNA-Seq data leading to observed noise and dropouts by constraining the latent dimension to $d << p$ where p is it total feature/gene space and d is the latent space. Since DCA estimates the mean expression parameter μ for each gene for each cell, we have used library size normalization on the μ parameters for both NB and ZINB models effectively generating 2 normalized datasets with imputed and non-imputed pre-processing.

C.0.1.3 Dimension Reduction

UMAP:Uniform Manifold Approximation and Projection (UMAP) is a nonlinear dimension reduction technique heavily utilized in scRNA-Seq data analysis. For a detailed explanation see the original article and python package documentation [31, 169]. Simply UMAP aims to build a k-nearest neighbor graph defined by the parameter 'number of neighbors' and generates an 'n' dimensional representation where distances of k-nearest neighbors are preserved across the dataset. Where possible, we have used default parameters 30 for 'n_neighbors' and 0.5 for 'min_dist'. For datasets with low number of cells we have set 'n_neighbors' as 10.

t-SNE: Similar to UMAP, t-Distributed Stochastic Neighbor Embedding (t-SNE) aims to find a low dimensional representation by minimizing KL-Divergence between pairwise distances in original space with low dimensional representation [30]. The perplexity parameter

is set to 30 for datasets with high number of cells 10 otherwise. θ parameter is set to 0.01 as well.

UMAP+PAGA: partitioned-based Graph Abstraction (PAGA) generates a low dimensional k-nearest neighbor graph embedding utilizing cell-cell similarities. Using PAGA embedded coordinates, UMAP can then be initialized with the PAGA coordinates allowing to couple PAGA with UMAP aiming to increase representative power of reduced dimensions [117].

DM: Diffusion Maps is a non-linear dimension reduction technique aiming to improve Principle Component based methods by incorporating 'diffusion' through construction of transition matrices. Transition matrices T_{nxn} are scaled distances representing affinities between pairwise cells converted to probability space by normalizing based on the sum of affinities of an individual cell. More specifically $T_{i,j}$ represents the probability of reaching cell j starting from cell i. Top n eigenvectors of the constructed transition matrix generates the reduced subspace [119, 170, 171].

Variational Autoencoder: Variational Autoencoders (VAE) widely used as generative models where the encoder model embedds the parameters of data prior and the decoder model generates the data from points sampled from embedded distribution. More simply, the encoder network embedds the mean and dispersion parameters where the prior distribution is assumed to be isotropic gaussian with mean 0 and standard-deviation 1 acting as a regularization. The neural network is constructed to have 3 hidden layers with $1024, 512, 256$ units for both the encoder and decoder network and a single stochastic latent layer with 2 units. We have used RMSprop optimizer with a learning rate of 0.0001

C.0.1.4 Clustering

Leiden: is an unsupervised clustering algorithm commonly used for scRNA-Seq data analysis. Modularity is used as the objective function to optimize and defined as the difference between the actual number of edges and the expected number of edges in a given

cluster. More specifically modularity is defined as $H = \frac{1}{2m}\sum_c(e_c - \gamma\frac{K_c^2}{2m})$ where γ is a resolution parameter positively associated with the number of clusters (threshold to define a cluster) [116].

tooManyCells: is an unsupervised method for simultaneous clustering and visualization of quantitative data. tooManyCells utilizes an efficient spectral clustering schema to recursively bipartition the cell-cell similarity matrix by using Newman-Girvan modularity as a stopping criteria [120].

C.0.1.5 Trajectory Mapping

Slingshot: utilizes principle curves to generate cell orderings given the minimum. Briefly, Slingshot uses initial clustering of cells followed by minimum-spanning tree (MST) identification of the cell clusters. Identified MST structure generates a cluster lineage which is followed by principle curve fitting simultaneously for each branch in the MST [121].

Palantir: utilizes nearest neighbor graphs extensively where diffusion maps are used to generate a denoised low dimensional representation of preprocessed scRNA-Seq data from an initial nearest-neighbor graph. Pseudotime estimates are then defined relative to an early cell (user defined) as the shortest path from the early cell. Pseudotime estimates are weighted based on waypoint cells iteratively sampled from the top n diffusion components [122].

DDRTree: aims to find a regularized low dimensional projection ($d << D$) of original high-dimensional dataset onto a space formed by orthogonal set of basis vectors with minimum-spanning tree (MST) regularization term [172, 173]. Also note that due to computational constraints, we have initially reduced the data to 50 principal components. Default values are used for remaining set of parameters.

Waddington-OT: is a supervised trajectory mapping method where sequencing time-points are used to construct transition probabilities between consecutive cell populations. Transition probabilities are estimated by optimizing an unbalanced optimal-transport problem

where the aim is to find a mapping between cells minimizing the cost incurred by transcriptional mass difference (euclidean distance of transcriptional profiles) [124].

C.0.1.6 Trajectory Comparison

We have extensively utilized Spearman's ρ for comparison of PTEs, and where there are multiple trajectories for a dataset, entropy as a measure to quantify overlap of PTEs obtained from different workflows. Specifically for Slingshot PTEs, we used entropy to quantify the distribution of spearman's ρ (scaled to $0-1$) between all pairwise PTEs from different workflows. Low entropy would suggest a better overlap where the pairwise PTEs would correspond to spearman's ρ $-1, 1$ Since Slingshot can generate > 1 trajectory, in order to evaluate the distribution of rank correlations between pairwise trajectory comparison, we calculated entropy to quantify whether Spearman correlations are bimodal around $0-1$ corresponding to 'high' quality overlap between trajectories (**Fig.**C5). DDRTree/Monocle2 trajectory comparisons however, are solely based on the pairwise geodesic distances of cells embedded on the trajectory tree identified by the method. This allowed us to both do evaluation unbiased to root node selection and amenable to automation.

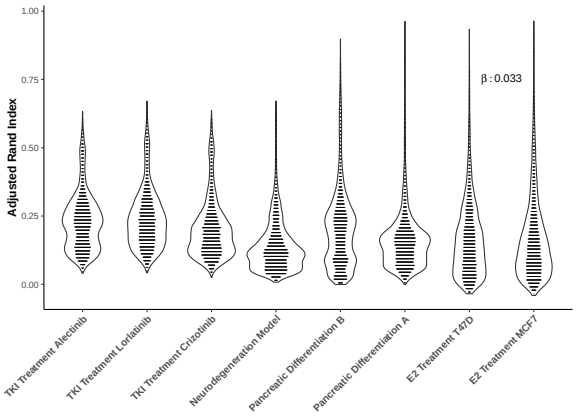
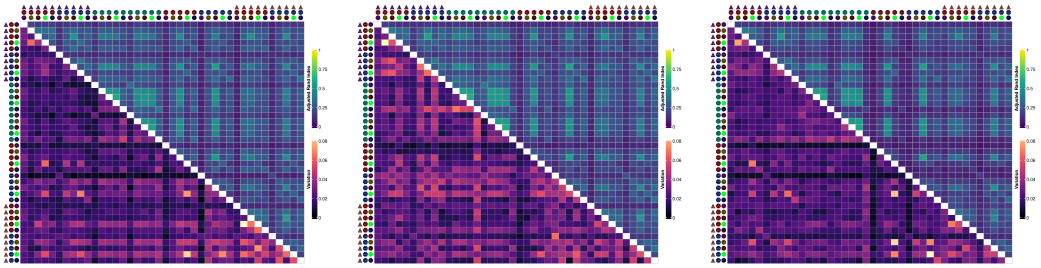


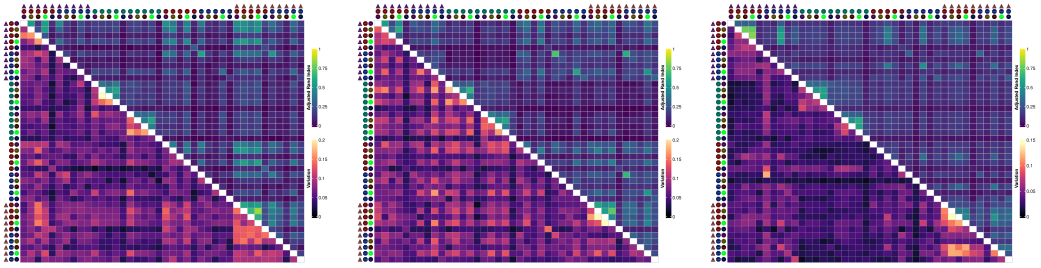
Figure C1: **Adjusted rand index distribution across different datasets ordered by decreasing number of cells.** Each point represents a pairwise comparison of clusters identified using different combinatorial workflows. Linear regression of ARI using number of cells as a covariate shows significant association with $\beta = 0.03$ ($p < 0.001$).



(a) NSCLC cell line Alectinib treatment

(b) NSCLC cell line Lorlatinib treatment

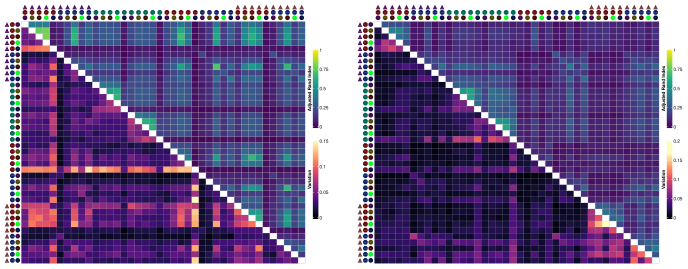
(c) NSCLC cell line Crizotinib treatment



(d) MCF7 cell line E2 treatment

(e) T47D cell line E2 treatment

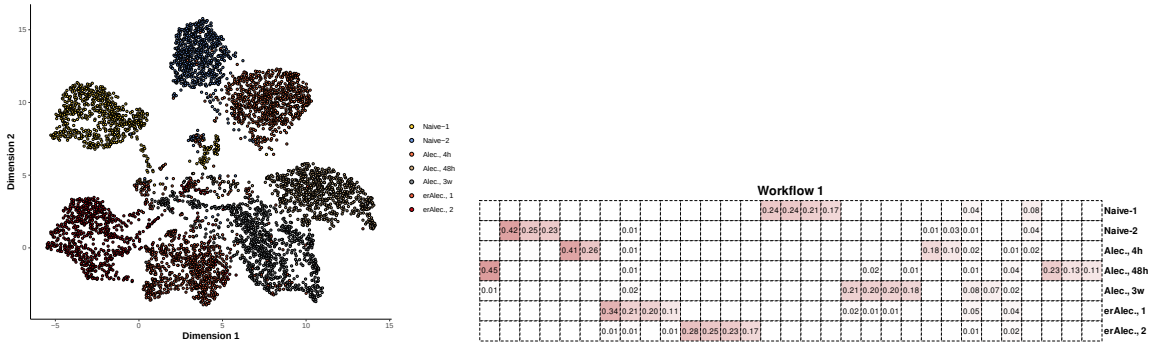
(f) Maturation of Pancreatic islets α cells



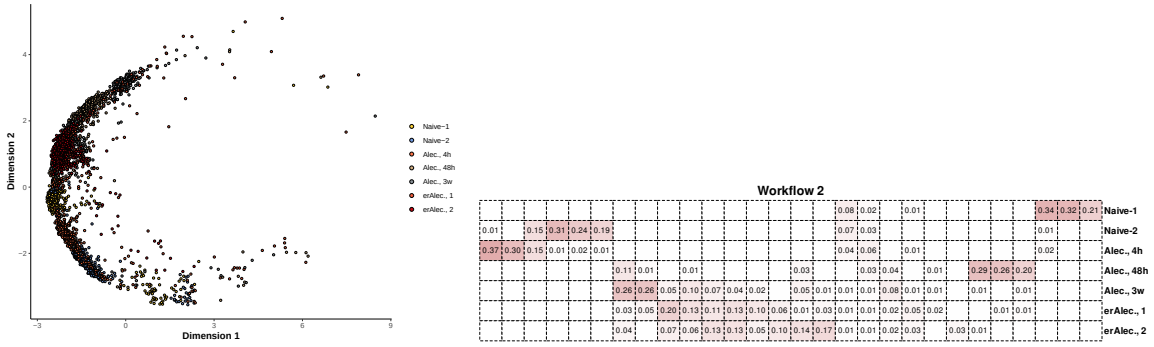
(g) Maturation of Pancreatic islets β cells

(h) Neurodegeneration mouse model

Figure C2: **Comparison of clusters identified using Leiden clustering.** Adjusted rank index across different methods is used to evaluate cluster overlaps which is further summarized by calculating the median across 12 subsets. Combinatorial workflows are also represented with icons depicting different levels of analysis steps. (a-c) TKI Treatment dataset, (d,e) E2 treatment dataset, (f,g) Pancreatic islet cell maturation and (h) Neurodegeneration dataset Overall t-SNE shows cluster identification which are relatively robust to different preprocessing steps.



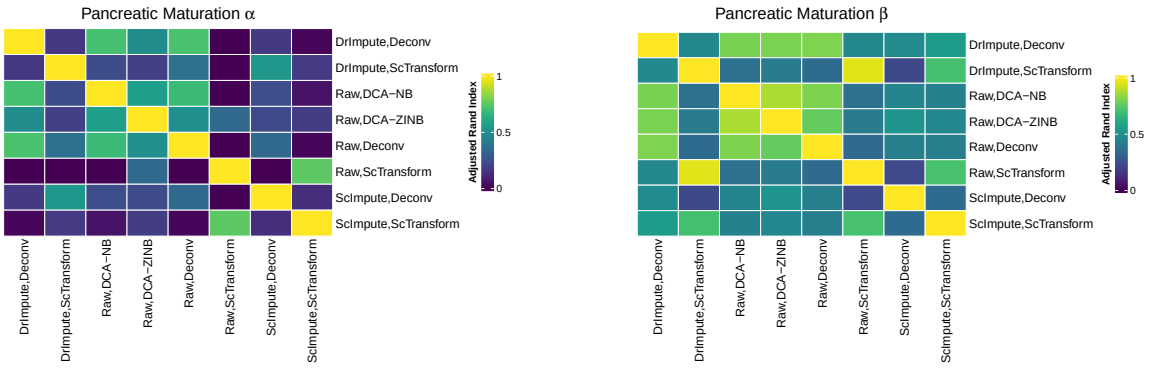
(a) DCA(NB),UMAP+PAGA



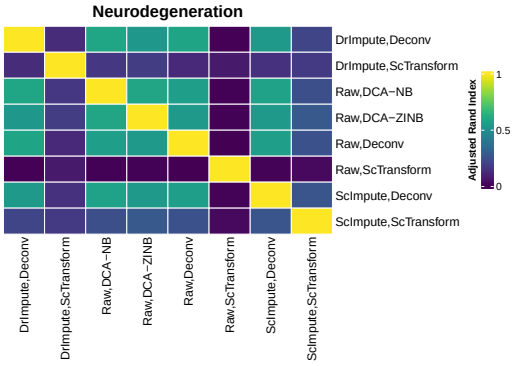
(b) Deconvolution,VAE

Figure C3: **Example cluster identification showing distinct results for both reduced dimensional representation and identified clusters between UMAP+PAGA and VAE dimension reduction methods stressing the importance of method selection for scRNA-Seq clustering analysis.**

Heatmaps show the row-normalized percentage of cells at specific sampling time-point (rows) assigned to distinct clusters (columns) in 1 data subset out of 12.



(a)



(b)

Figure C4: **tooManyCells cluster overlap quantified by ARI showing relatively good overlap in the Pancreatic Maturation dataset.** However, data-specific performance of different steps are present where α cells and β cells datasets show opposing trends in combination of imputation and normalization.

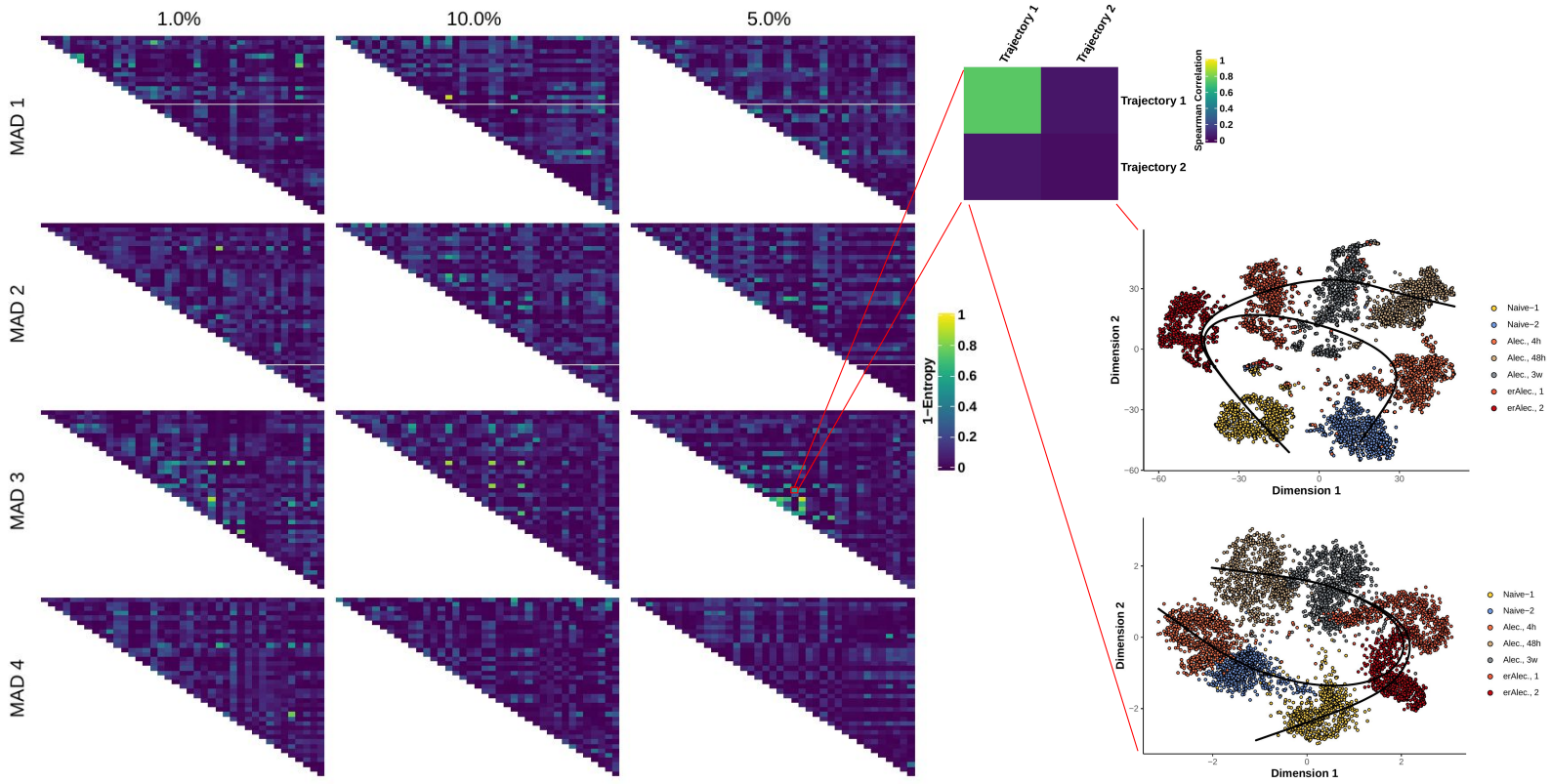
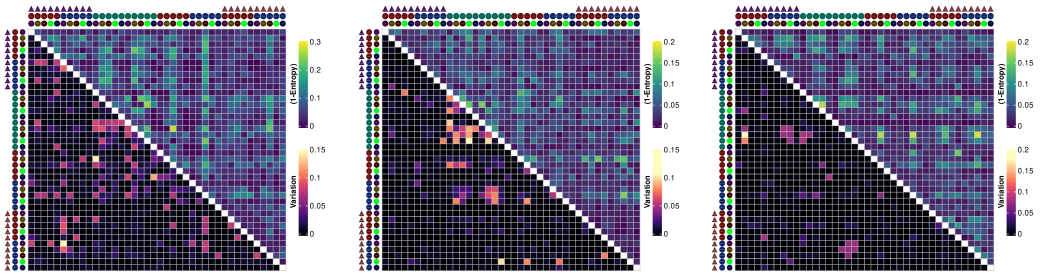


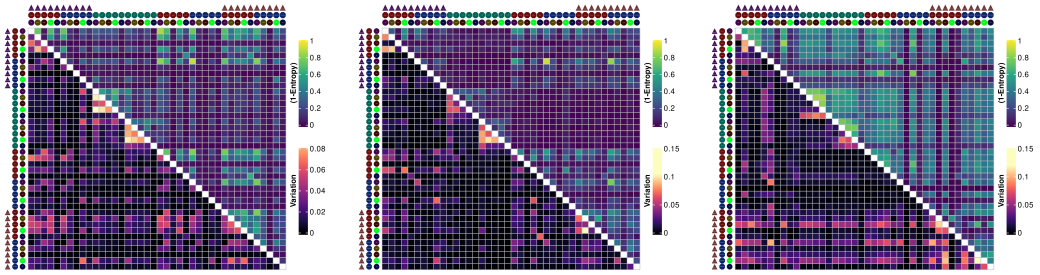
Figure C5: **Example figure showing the quantification by entropy over multiple trajectories identified** In order to quantify the global overlap we have compared inidividual trajectories using spearman correlation scaled between 0-1. Using the scaled spearman correlations as pseudo-probabilities, we calculated entropy to assess whether scaled correlations are centered around 0-1 suggesting good overlap.More specifically, if the correlations are centered around 0-1, this suggests a bimodal mapping of identified trajectories across different workflows hence 'good' overlap. In contrast, if the rank correlations are distributed around 0.5, the overlap of trajectories are mostly random and an individual trajectory can map to multiple trajectories in the compared workflow.



(a) NSCLC cell line Alectinib treatment

(b) NSCLC cell line Lorlatinib treatment

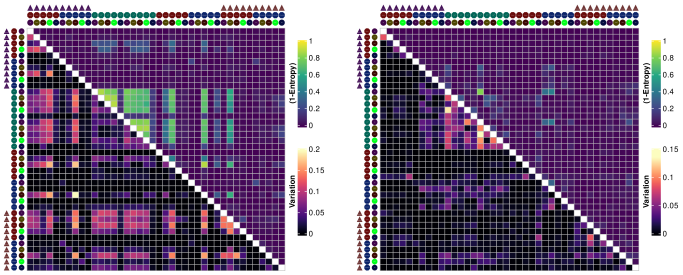
(c) NSCLC cell line Crizotinib treatment



(d) MCF7 cell line E2 treatment

(e) T47D cell line E2 treatment

(f) Maturation of Pancreatic islets α cells



(g) Maturation of Pancreatic islets β cells

(h) Neurodegeneration mouse model

Figure C6: **Comparison of trajectories identified by Slingshot.** Quality of overlap is summarized by quantifying the 'randomness' of scaled Spearman rank coefficients between the trajectories. Treating scaled rank coefficients as pseudo-probabities and using entropy allowed us to assess whether the pairwise trajectory comparisons are bimodal around 0 and 1 (suggesting good mapping/low entropy) or uniformly distributed (suggesting no optimal mapping). 1-Entropy values are then averaged across 12 subsets. Upper triangle shows the aggregated entropy values and lower triangle shows the variation in entropy values (Best overlap would be represented by low entropy and low variation values).

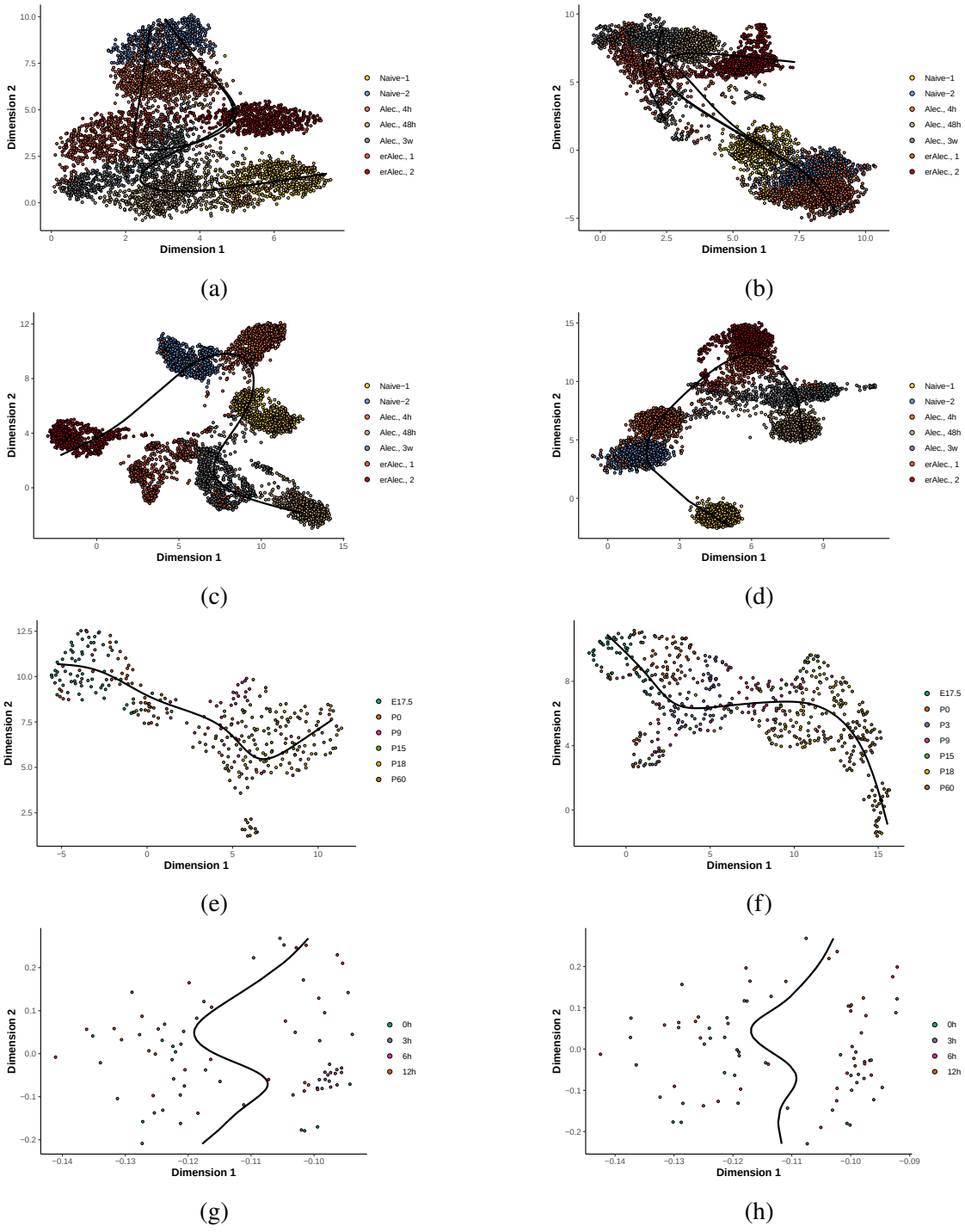
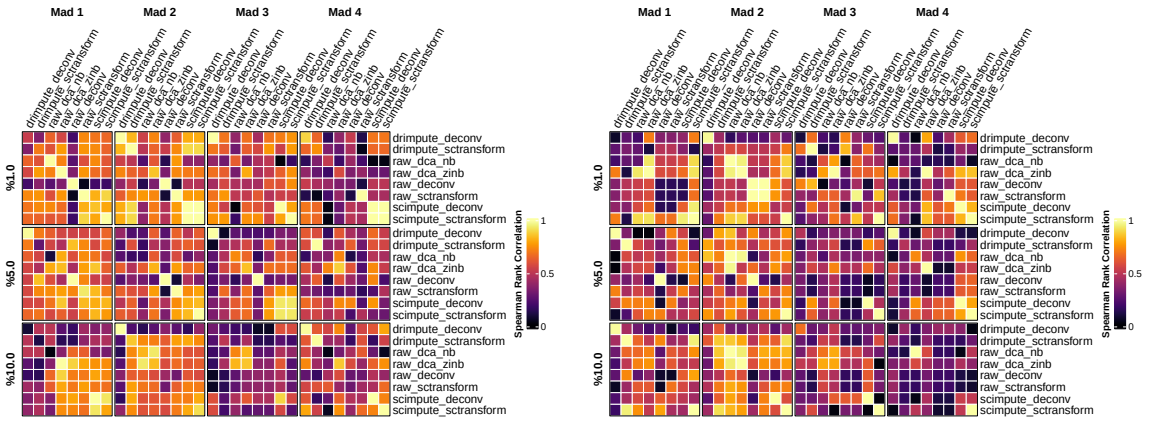
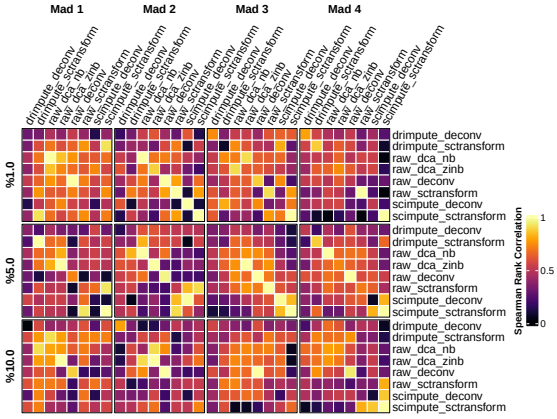


Figure C7: **Example Slingshot trajectory estimates** using (a-b) Deconvolution and ScTransform coupled with PAGA+UMAP on non-imputed dataset (c-d) Deconvolution and ScTransform coupled with PAGA+UMAP on imputed data with DrImpute. (e-f) shows Slingshot applied on Pancreatic maturation α and β cells respectively processed using DCA and dimension reduced with UMAP+PAGA showing relatively good overlap. (g-h) shows DM applied in E2 treatment dataset for DrImputed and no-imputation respectively.



(a) Alectinib

(b) Lorlatinib



(c) Crizotinib

Figure C8: **Pseudotime comparison using Palantir in the TKI Treatment dataset.** Results for each TKI is shown separately for each of the 12 data subsets with increasing gene and cell level thresholds. Median aggregated spearman's ρ over 10 replicates is given.

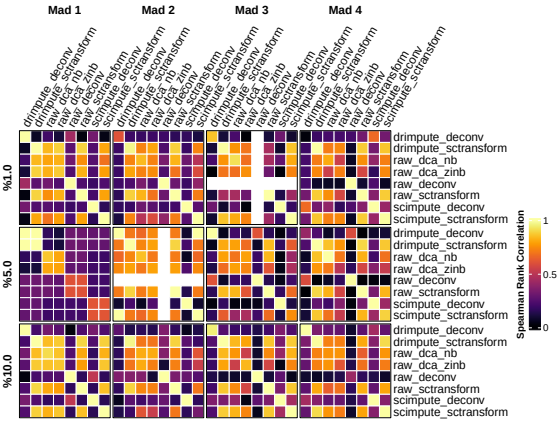
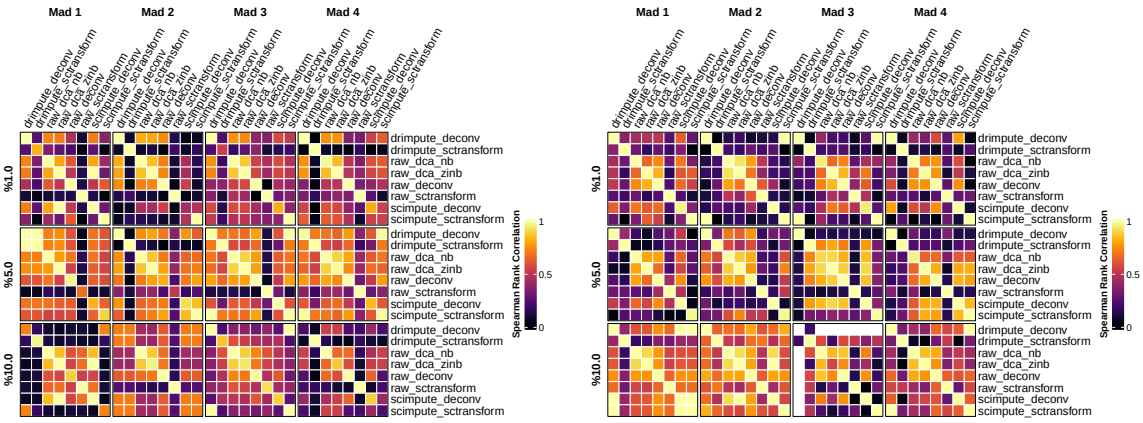


Figure C9: **Palantir pseudotime estimates in the Neurodegeneration dataset.**



(a) α cells (b) β cells

Figure C10: **Palantir pseudotime estimates in the Pancreatic Maturation dataset.**

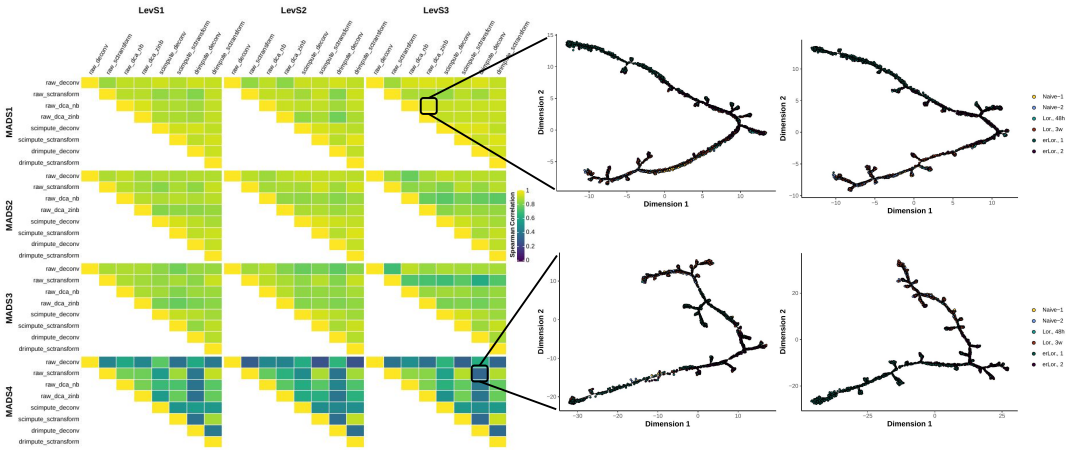


Figure C11: **Comparison of pairwise trajectories for Lorlatinib treated NSCLC cell line separately for 12 subsets.** Individual rank correlations are then aggregated by taking the median to summarize overall similarity of pairwise workflows.

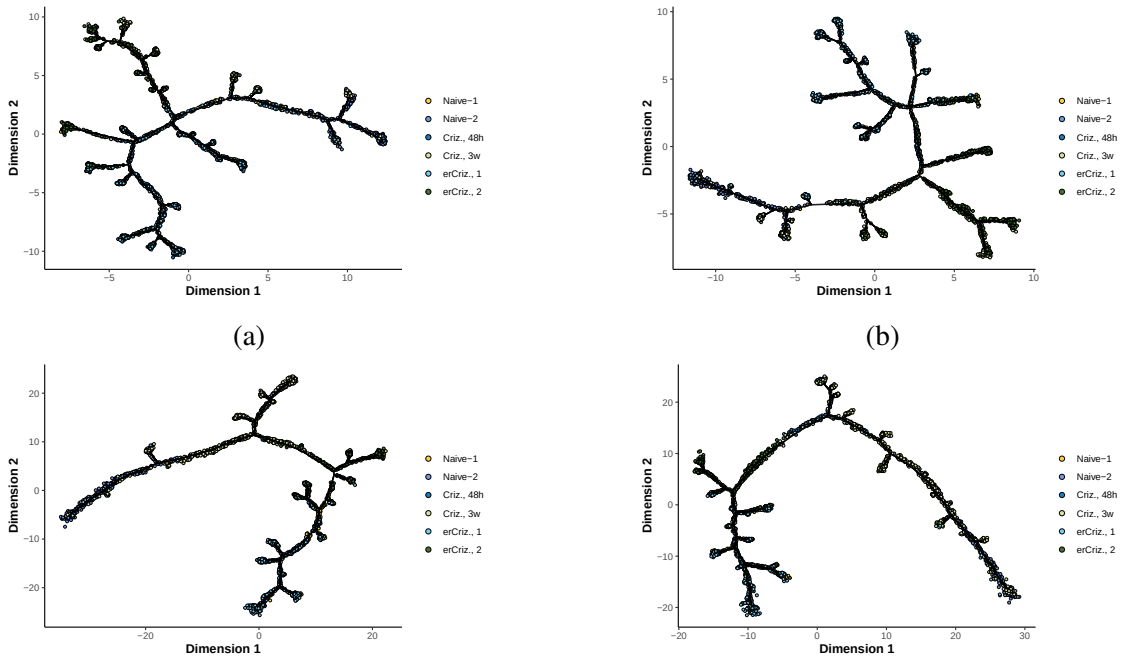
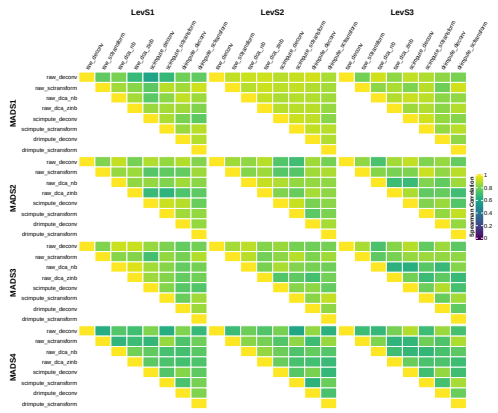
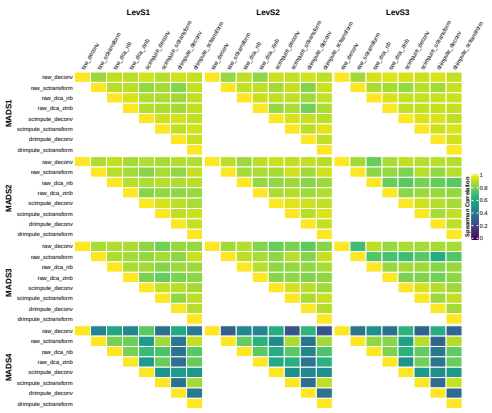


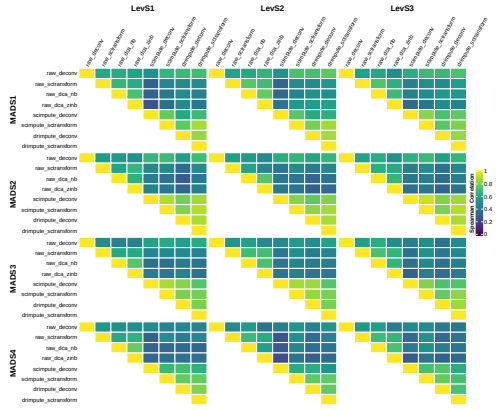
Figure C12: **Trajectories identified by DDRTree using Crizotinib dataset showing increased number of branch-points identified when DCA is utilized** (a-b) shows DCA-NB and DCA-ZINB respectively, (c-d) shows applying DrImpute and ScImpute followed by ScTransform respectively



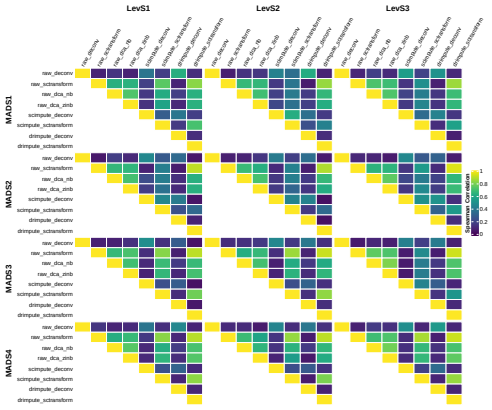
(a) Alectinib treated NSCLC



(b) Lorlatinib treated NSCLC

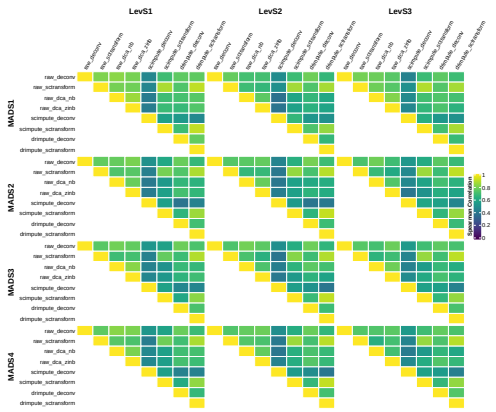


(c) Crizotinib treated NSCLC

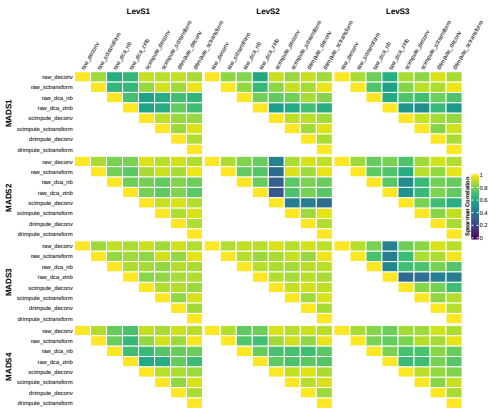


(d) Neurodegeneration model

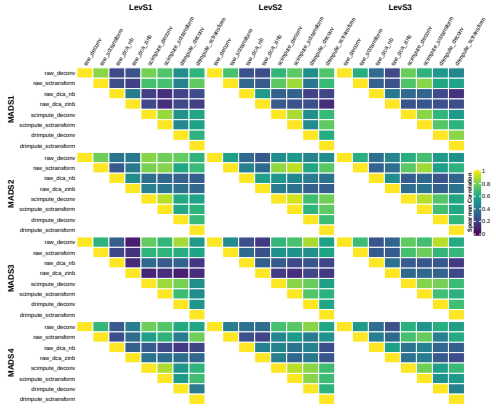
Figure C13: **Overlap of DDRTree trajectories across different subsets stratified by cell level (X-Axis), gene level (Y-Axis) thresholds and workflows quantified by the spearman rank correlation of geodesic distances between individual cells.** No substantial difference exists in rank correlations across different thresholds for pairwise workflow comparisons.



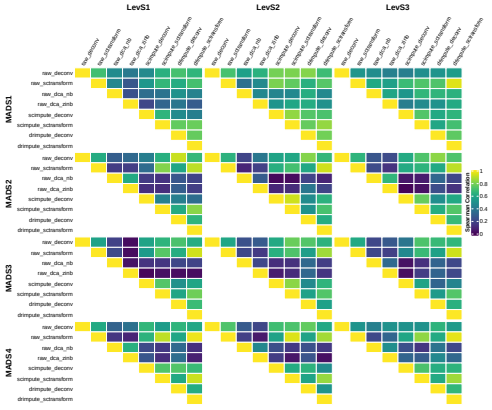
(e) Pancreatic differentiation α



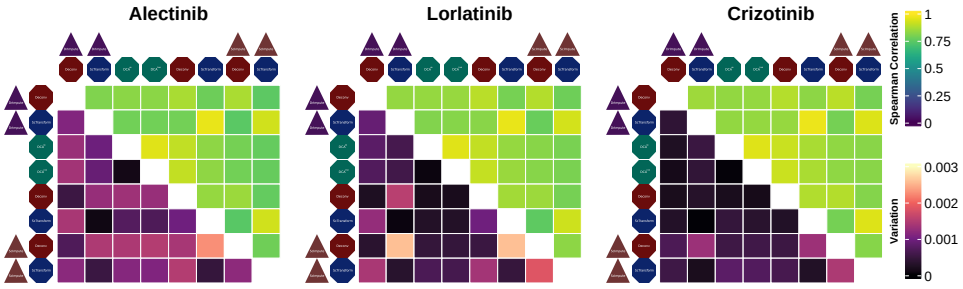
(f) Pancreatic differentiation β



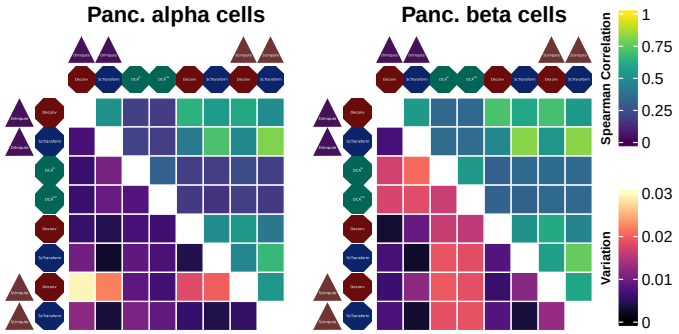
(g) E2 Treatment MCF7 cells



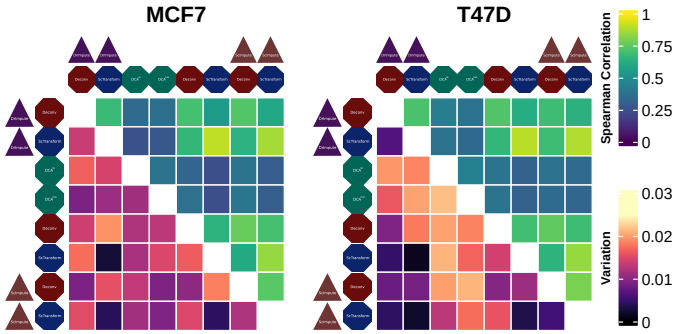
(h) E2 Treatment T47D cells



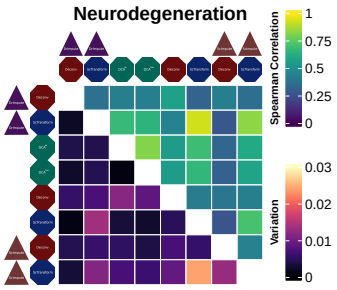
(a) TKI Treatment



(b) Pancreatic Differentiation



(c) E2 Treatment



(d) Neurodegeneration

Figure C14: **Waddington-OT PTEs comparisons showing median rank correlations across 12 subsets (upper-triangle) and associated variation (lower-triangle)**

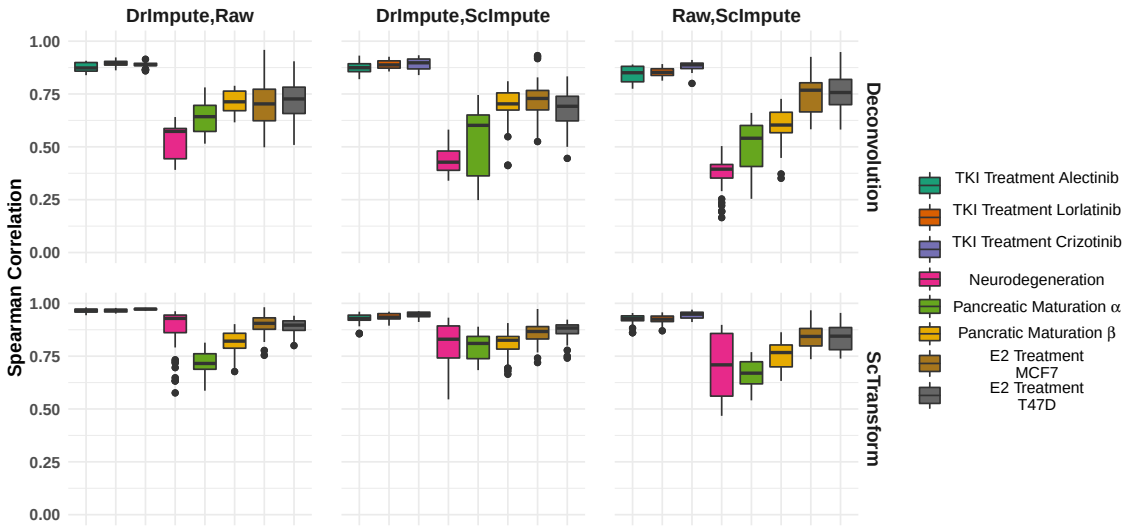


Figure C15: **Waddington-OT rank correlation comparison for normalization methods ScTransform and Deconvolution showing a global trend towards improved ScTransform PTE overlaps.** Individual points represent the rank correlation between different imputation workflows when Deconvolution and ScTransform is used as normalization step.

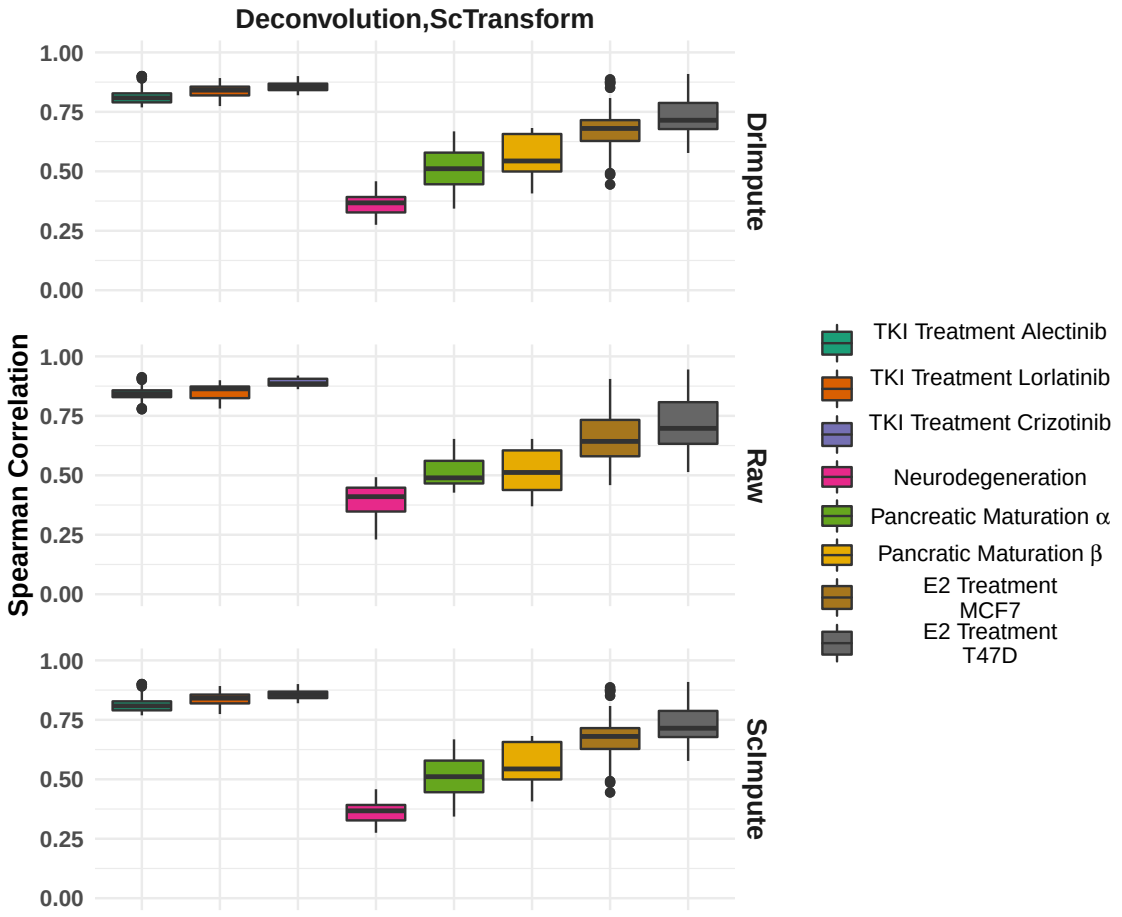


Figure C16: **Comparison of Imputation methods showing no substantial effect of preprocessing on WOT PTEs** Using ScTransform and Deconvolution for normalization shows no substantial dependence on imputation step hence resulting in similar rank correlations

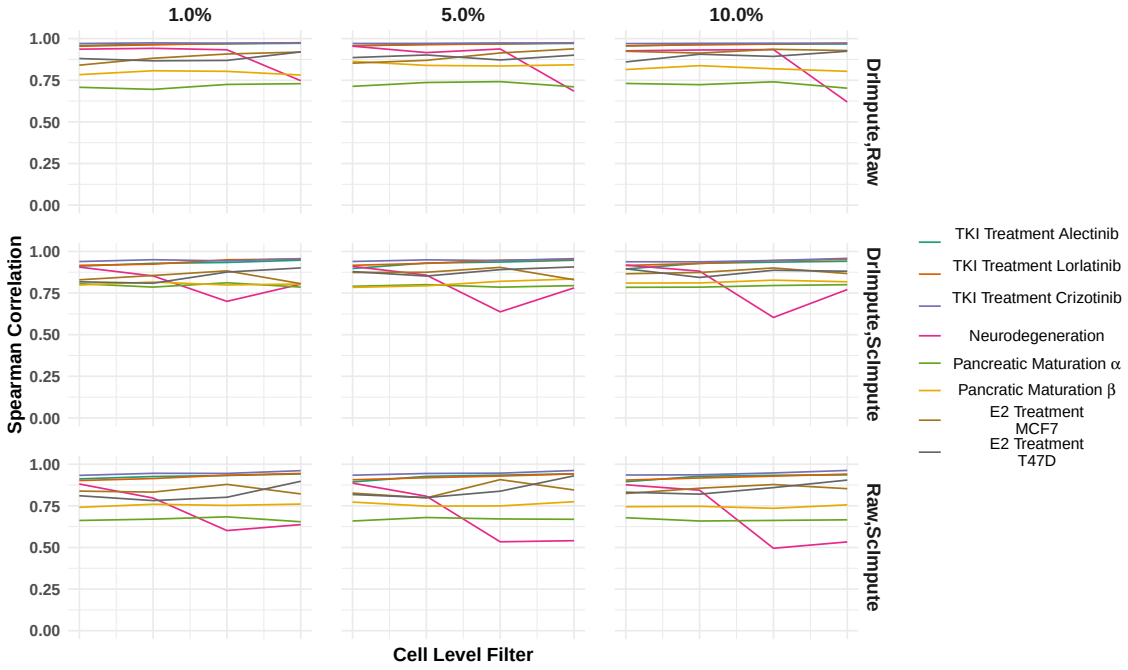


Figure C17: **WOT PTEs comparisons when ScTransform is used for normalization across 12 subsets separated by cell level and gene level filtering showing reduced effect.** Spearman rank correlations show no substantial difference when different thresholds are used for filtering out low quality cells and genes. x-axis is ordered in increasing cell level threshold and each facet is given in increasing order of gene level threshold (1%, 5%, 10%).

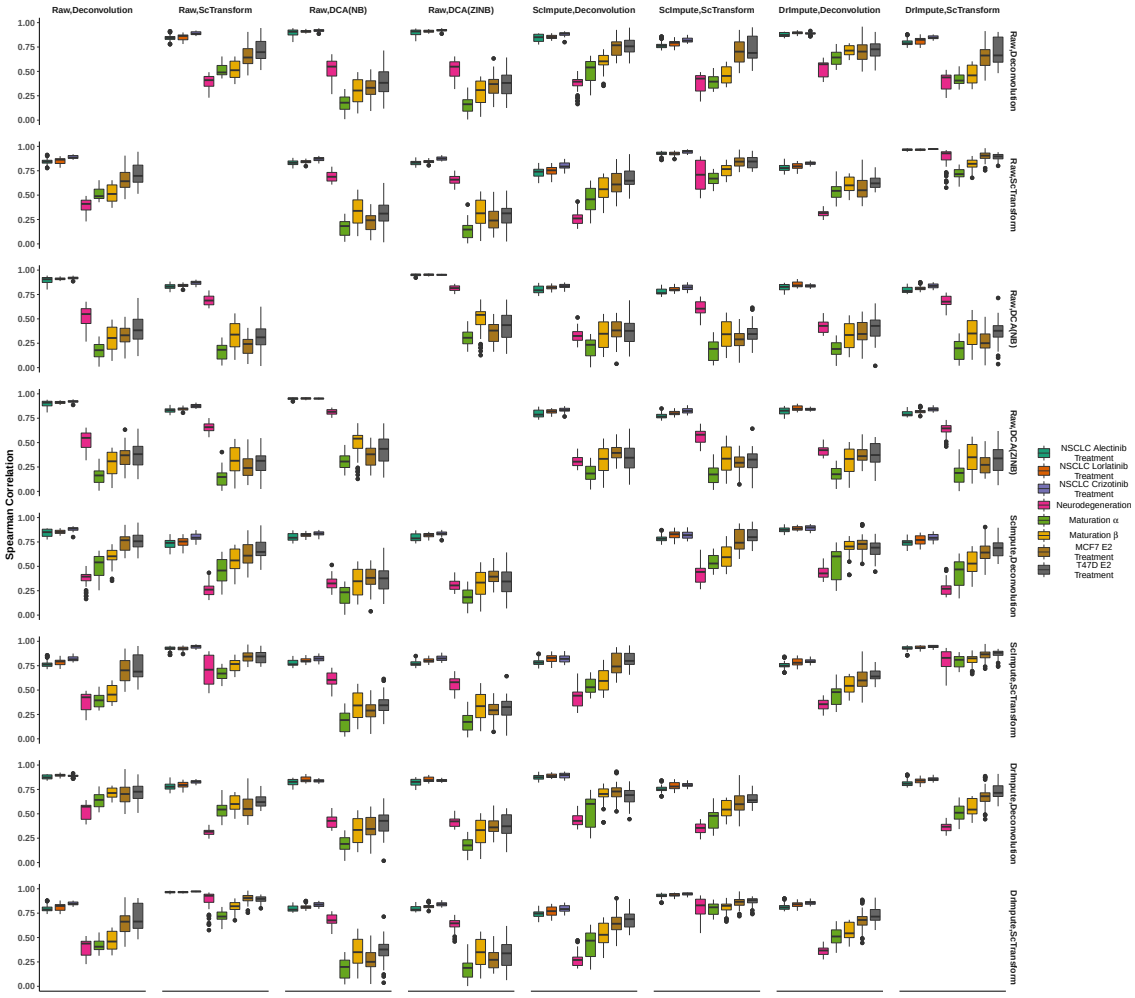


Figure C18: **Distribution of Waddington-OT rank correlations between different workflows.**

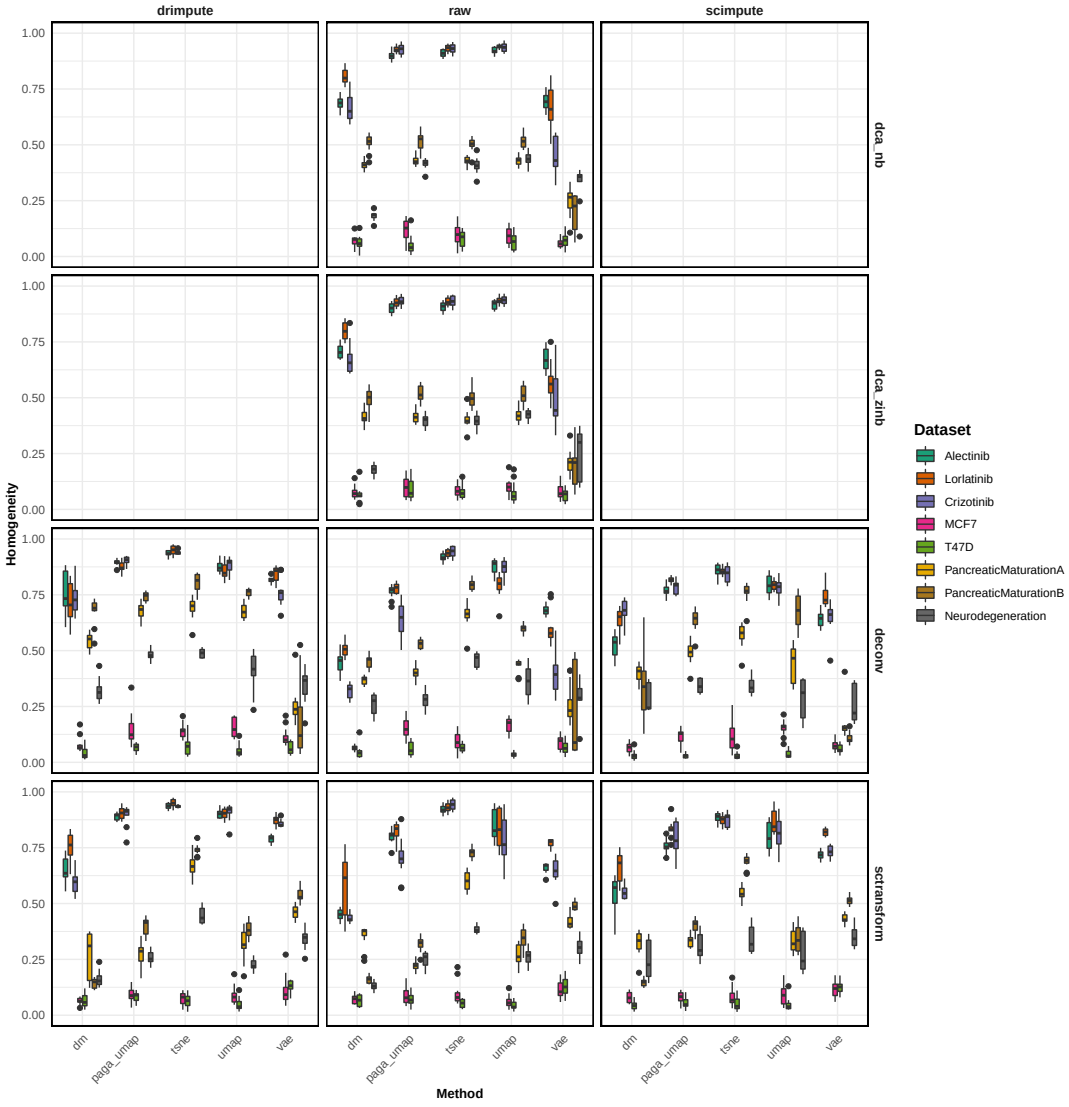


Figure C19: **Homogeneity of clustering in comparison to time-point labeling.** Specifically, homogeneity is quantified using normalized entropy where distribution of cells from different time-points in a single identified cluster decreased homogeneity.

Appendix D

INTEGRATIVE MODELING OF DRUG SENSITIVITIES USING MACHINE-LEARNING

Supplementary Methods

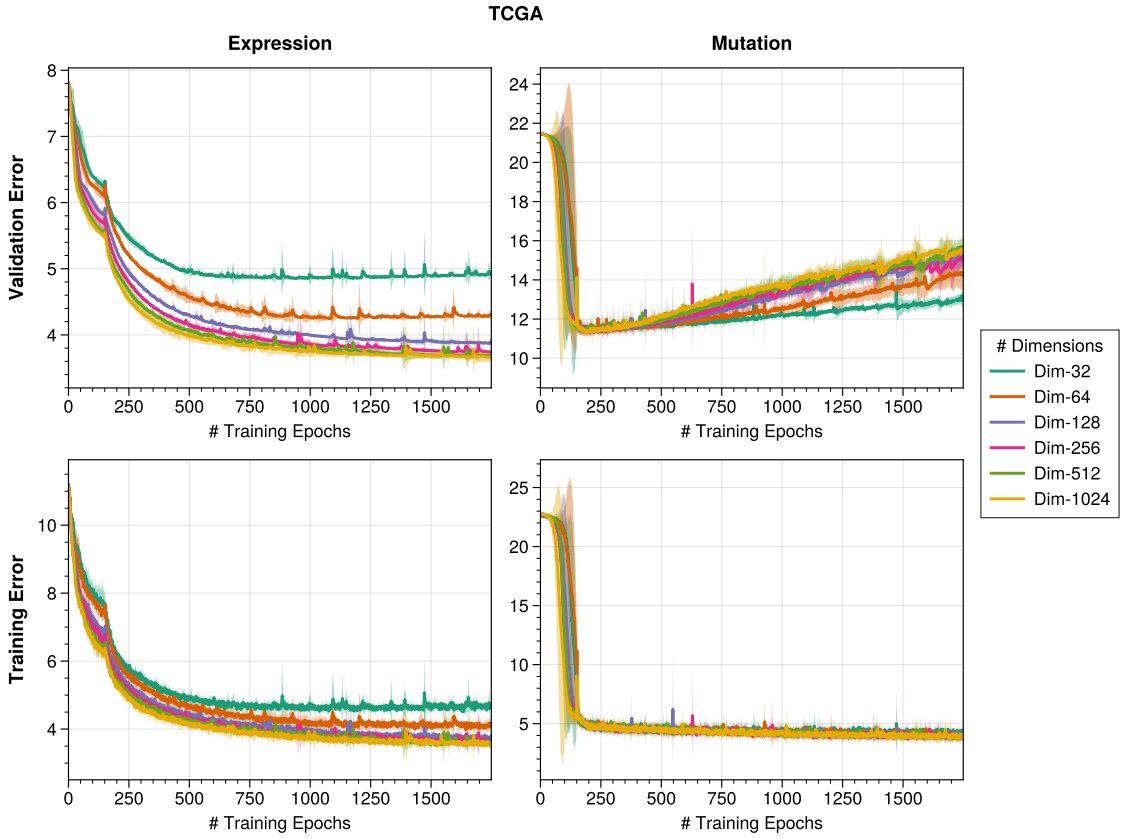
Autoencoder

In order to accurately learn mapping from -omic features to drug sensitivity, and to couple expression with mutation information, we have utilized an autoencoder neural-network architecture with gene-set regularization. More specifically, input expression and mutation profiles are first passed through a gene-set layer which is an p feature layer weighted by the mapping of genes to the given set (genes mapping to the given set are assigned a weight 1.0 and 0.1 if not). Following the gene-set layer, expression and mutation outputs are mean aggregated and further passed through a multiple bottleneck latent-feature layer supervised by drug-sensitivities, tissue-types and data-types (whether the data is a cell-line or patient sample) hence the network aims to learn to efficiently compress expression and mutation profiles accounting for relevant features associated with drug sensitivity. However, since the number of cell-lines is relatively low, we utilize TCGA dataset as well allowing the network to leverage relatively large patient dataset where a single batch node is also included at the final output layers to account for cell-line, patient sample batch effects. For the gene-set regularization layer, we utilized GO-Biological-Process ontology database keeping gene-sets with $10 < N_{genes} < 100$.

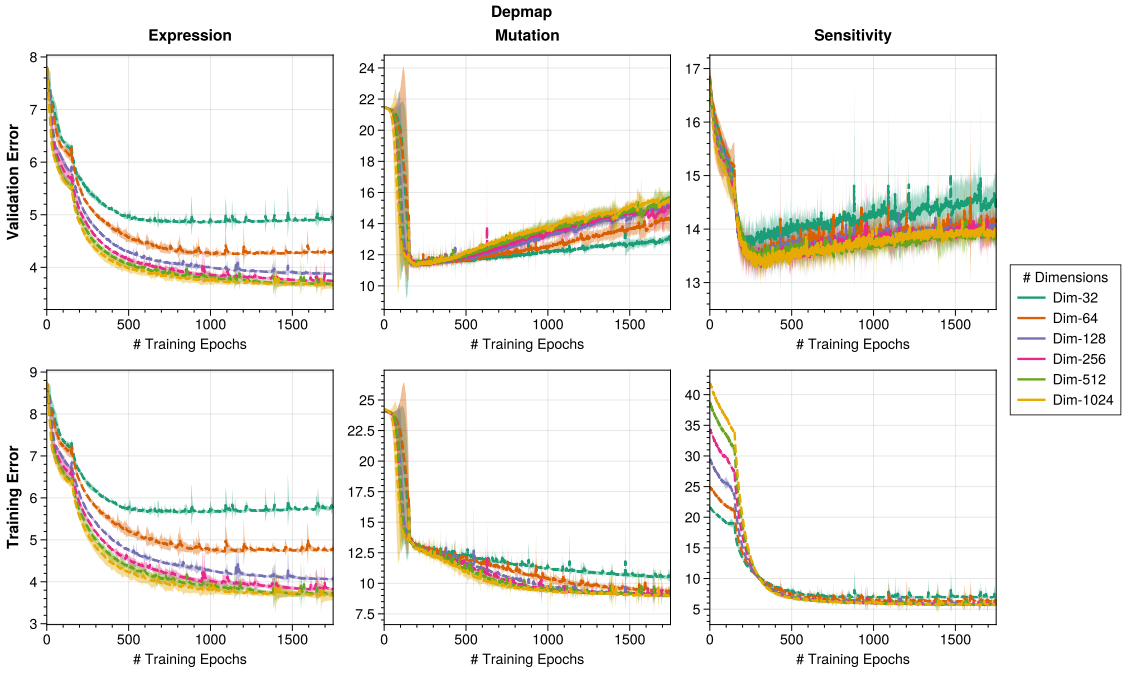
Preprocessing

We have restricted our analysis to protein-coding genes and further filtered to include genes with at least 20% of samples had > 12 reads mapped. Using variance stabilizing transformation in *DESeq2* package, we generated log-transformed expression values. Mutations aggregated into gene-level profiles are filtered to include only genes annotated as possible

drivers using the IntOGen database. Mutations are filtered to include $VAF > 0.05$.

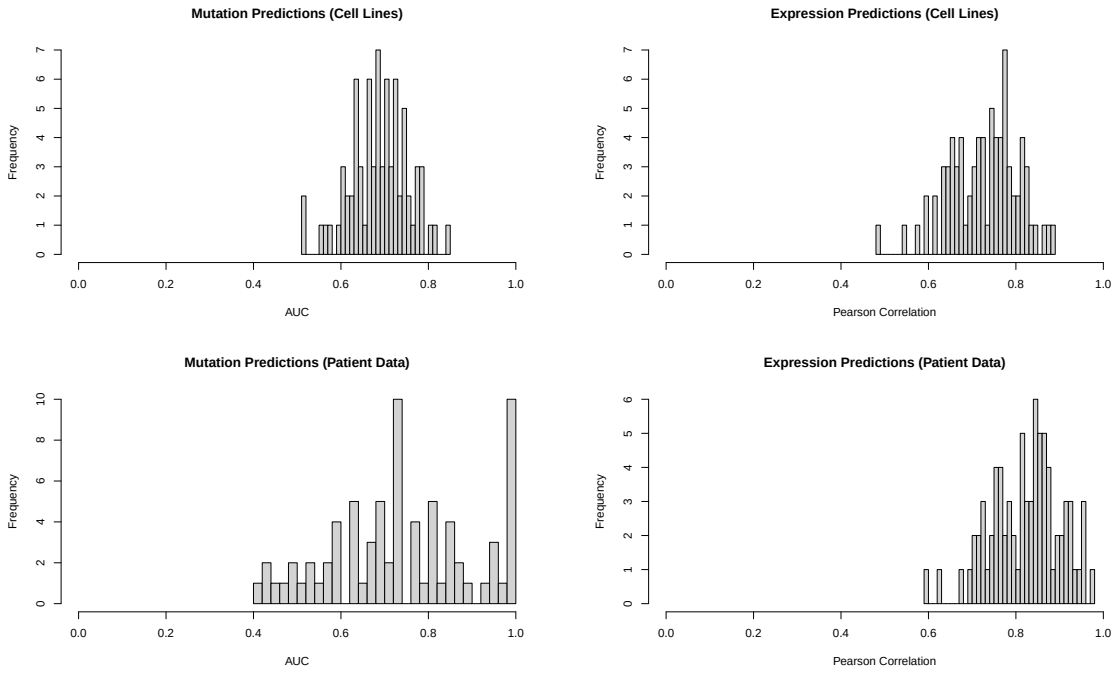


(a)



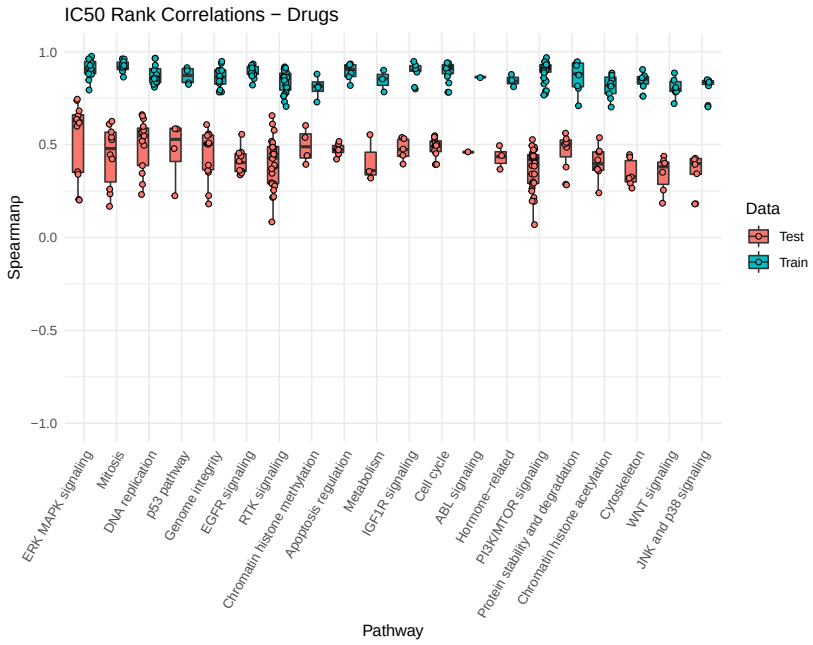
(b)

Figure D1: Cross-validation across multiple latent dimensions in (a) TCGA patient dataset and in (b) DepMap cell-line dataset. We have used mean-squared error summed over batch for gene expression, drug sensitivity and log-loss summed over batch for binary mutation profiles. Results shown are mean aggregate of 5 training runs.

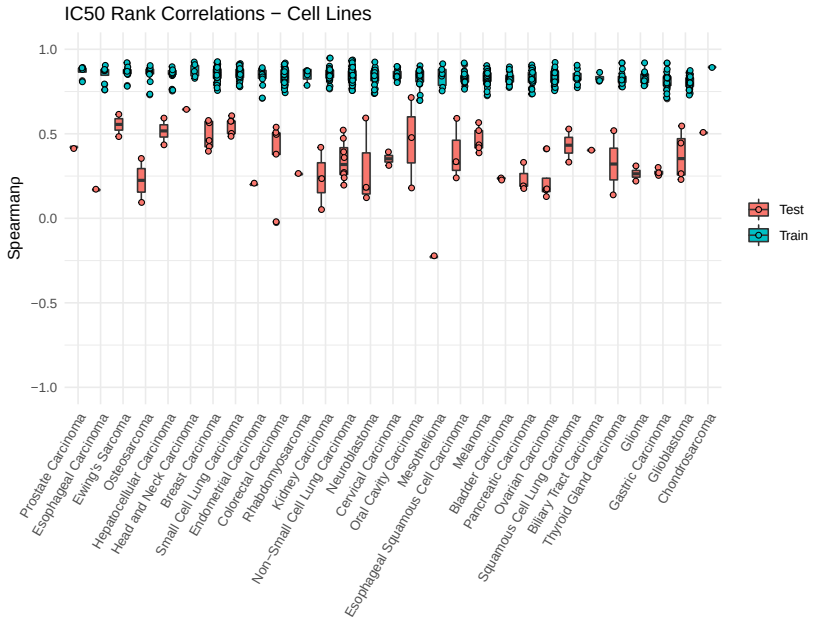


(a) (b)

Figure D2: Histogram of prediction performance in the test dataset measured by area under the reciever operator curve (AUC) (a) and pearson correlation (b)

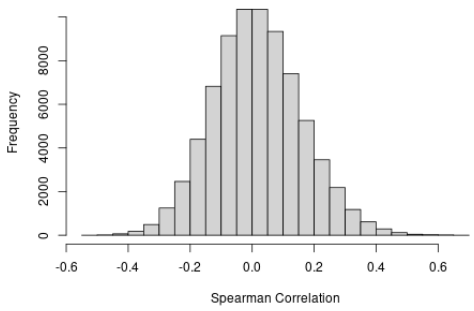


(a)

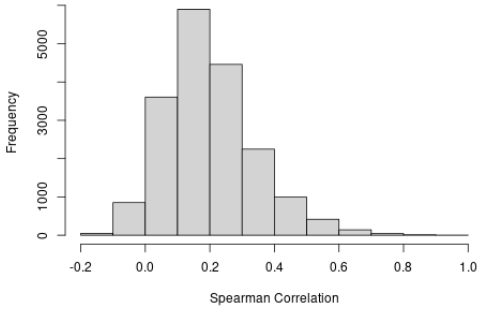


(b)

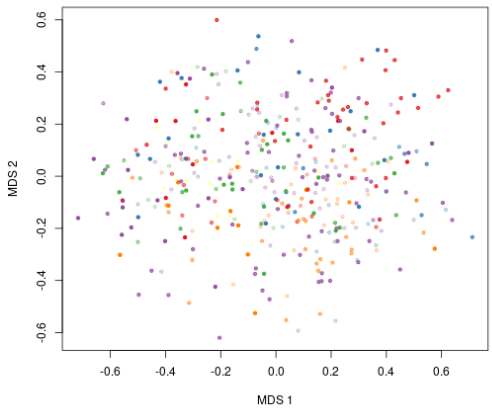
Figure D3: Overlap of model predictions quantified by Spearman's ρ stratified by target pathways and cancer types respectively for drugs and cell-lines showing relatively high overlap in the training data but reduced overlap in the test dataset.



(a)

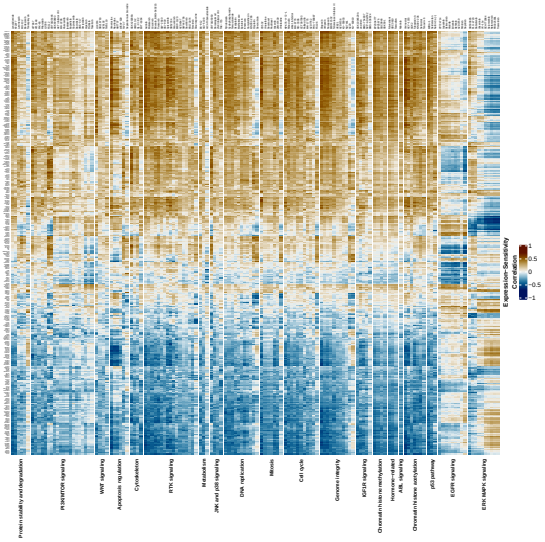


(b)

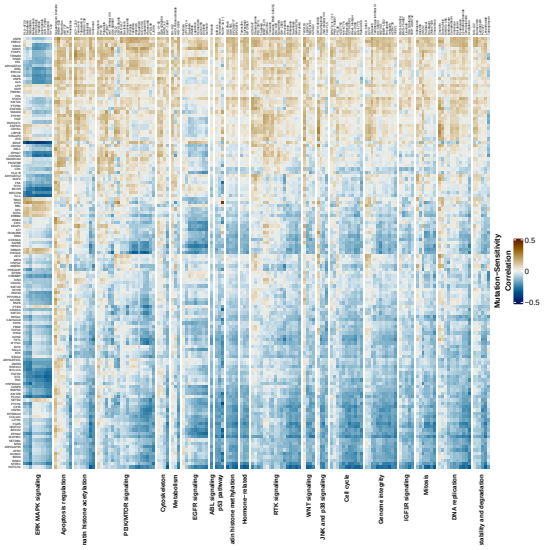


(c)

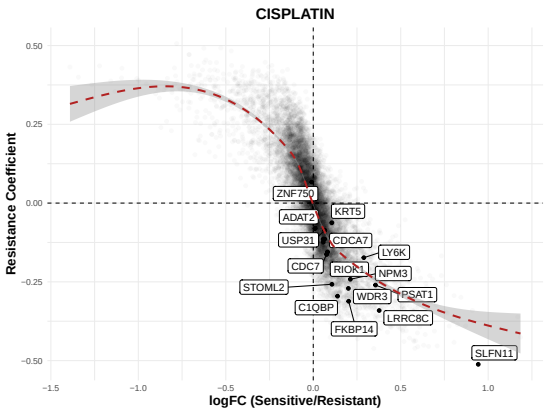
Figure D4: Similarities of drug sensitivities quantified by spearman correlation across cell-lines (a) and drugs (b) showing reduced linear associations when considering all the drugs and cancer types. Multidimensional scaling of cell-lines further demonstrating increased dispersion suggesting 'uniqueness' of drug responses.



(a)

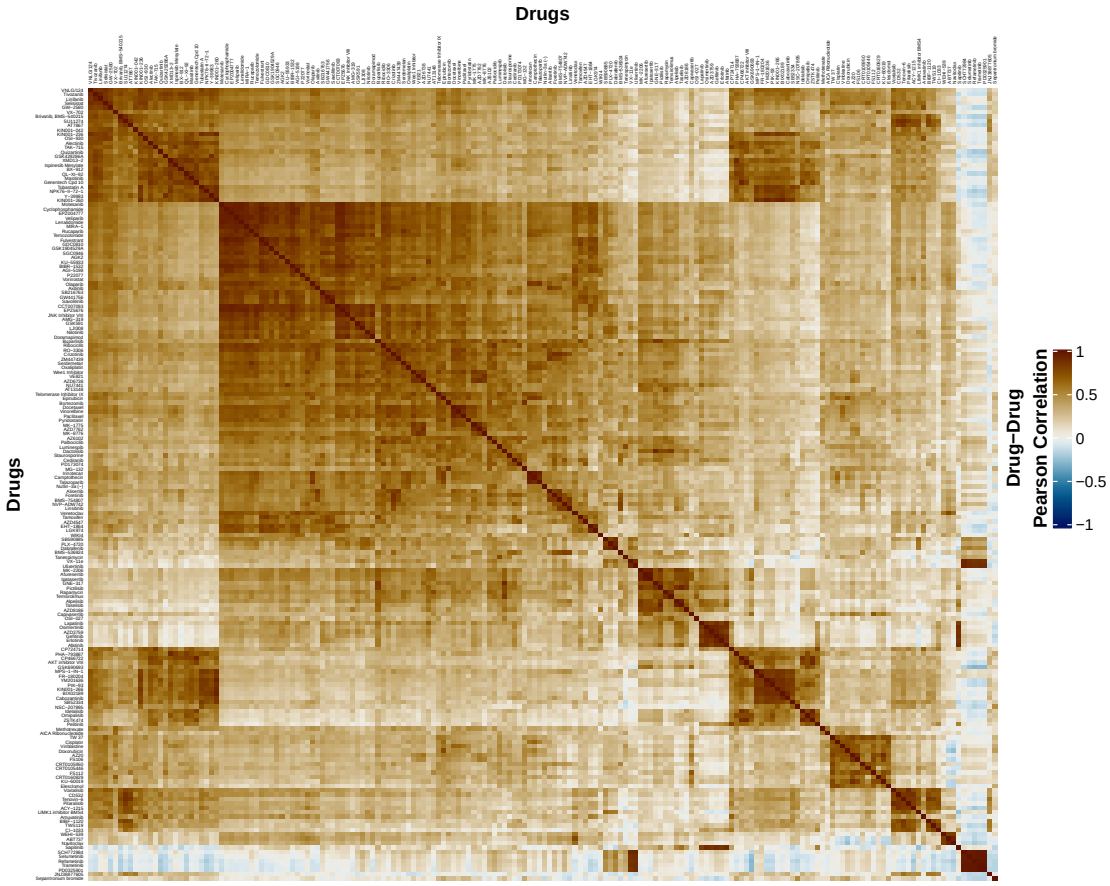


(b)



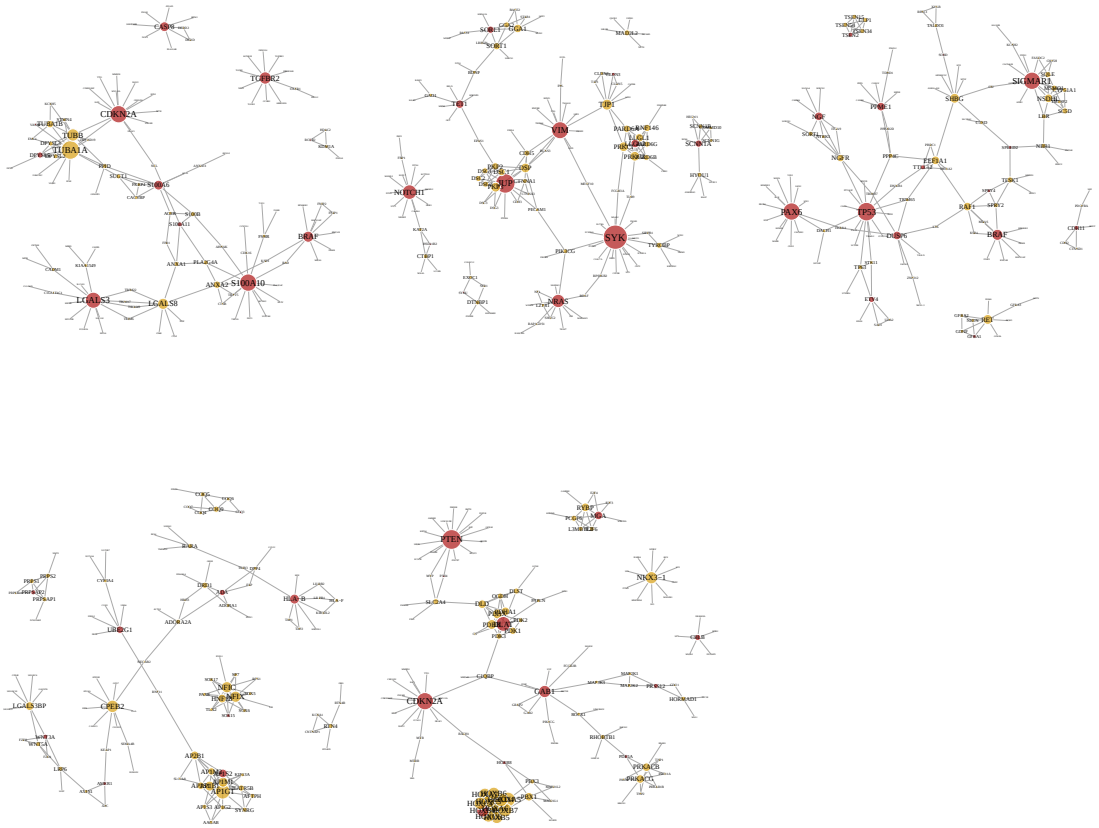
(c)

Figure D5: Multi-omic feature associations quantified by sampling the latent space and calculating Pearson correlation. Gene-expression associations showing top 10 signature genes for each drug (a). Mutation associations showing top 3 signature genes for each drug (b). Genes known to be positively associated with Cisplatin sensitivity overlapping with (-) resistance coefficients



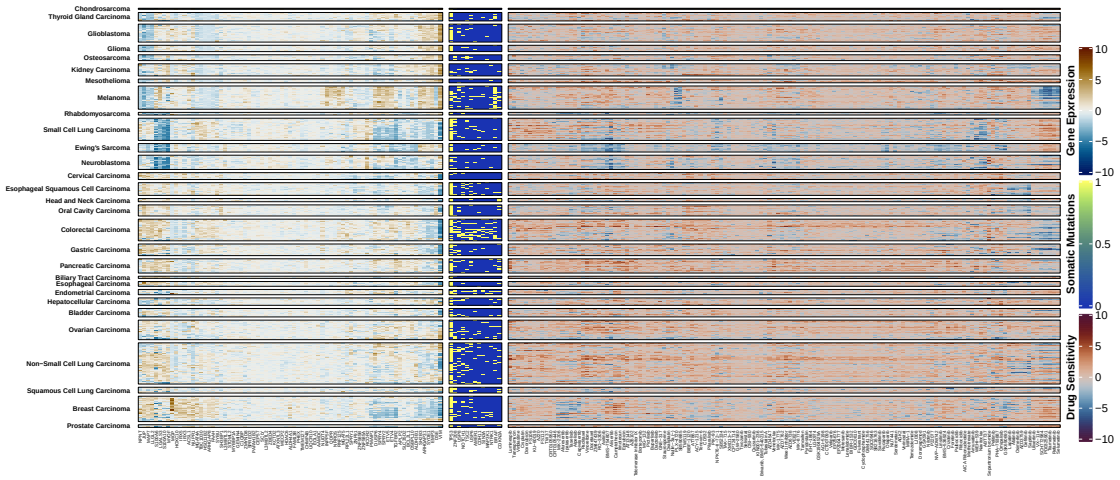
(a)

Figure D6



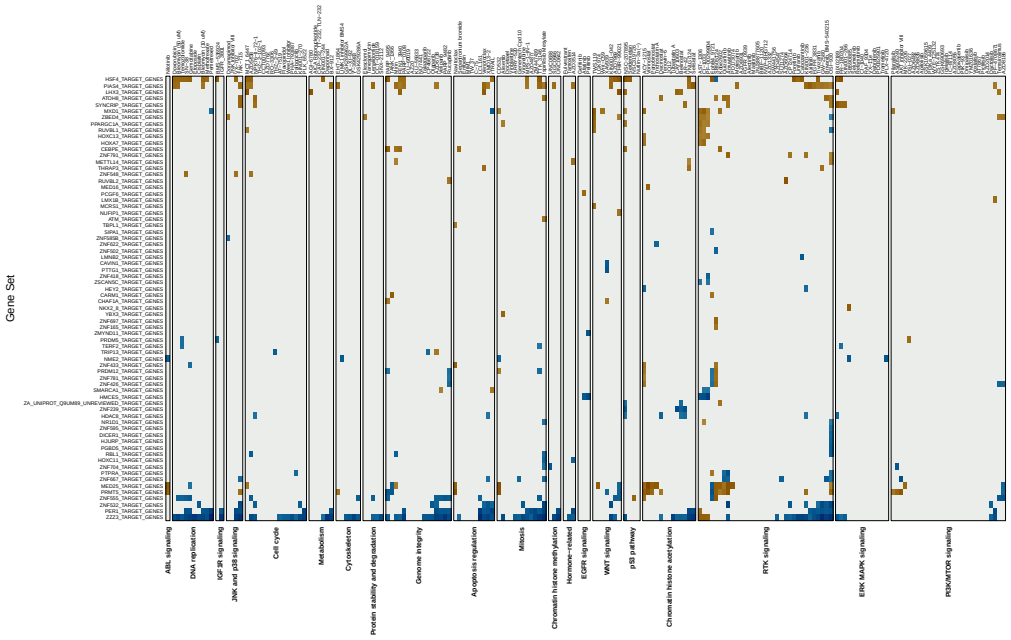
(a)

Figure D7: **Gene-gene interaction subnetworks identified through biased-random walks for top features with high 'loadings' on singular vectors obtained by svd on feture-drug correlation matrix for drugs with high-frequency of negative interactions**

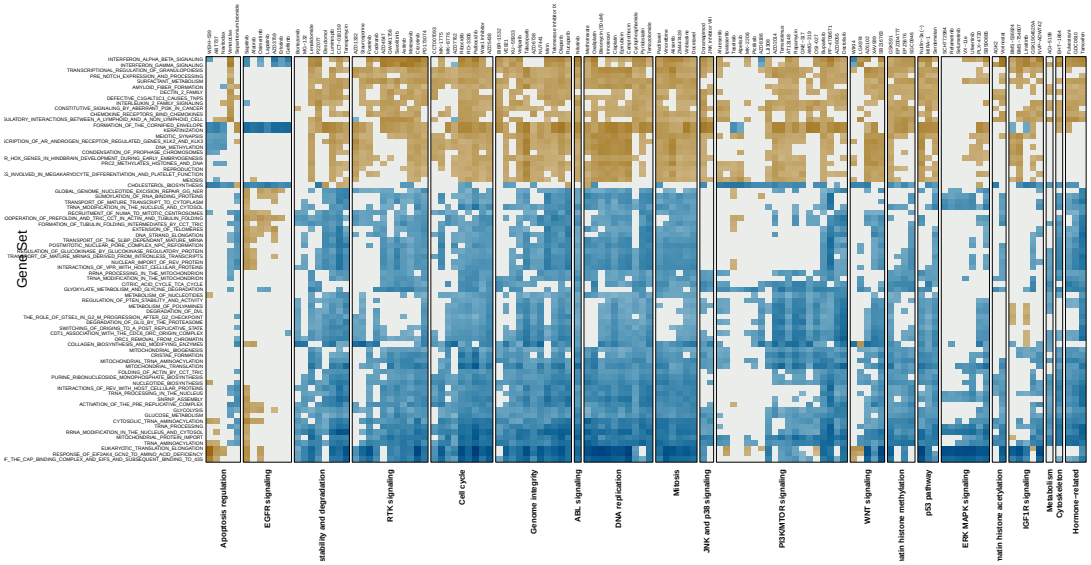


(a) **Omic profiles for top associated features across drugs** Heatmaps show both gene expression, mutation, and IC50 values for drugs that showed $\rho > 0.4$ predictions in the test set. Features are selected based on the 'loadings' on singular vectors of svd applied on feature-drug correlation matrix

Figure D8

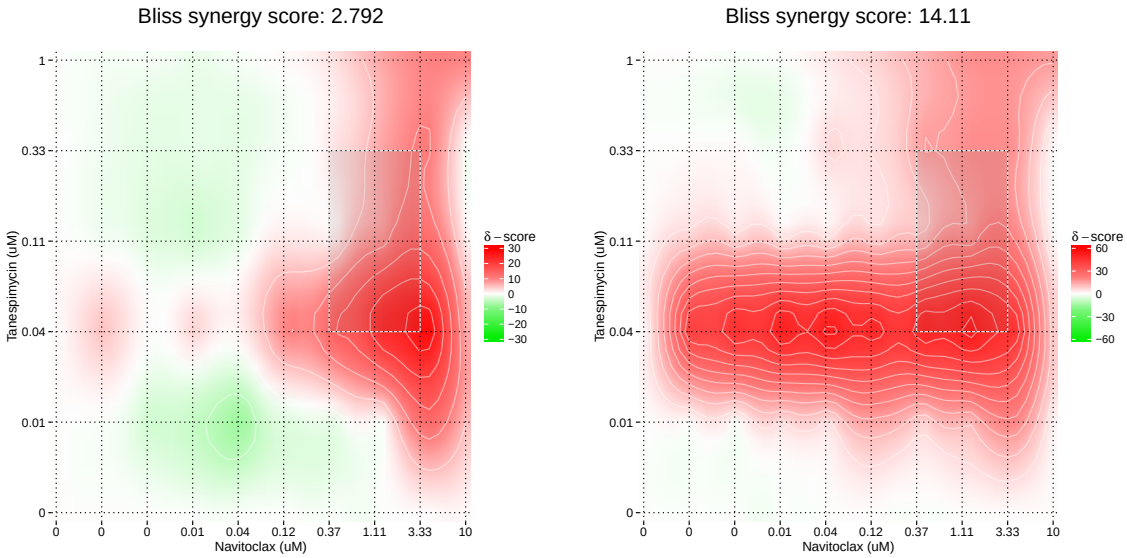

Gene Set
GSEA NES
4
2
0
−2
−4
ABL signaling
DNA replication
IGF IR signaling
JNK and p38 signaling
Cell cycle
Metabolism
Cytoskeleton
Protein stability and degradation
Genome integrity
Apoptosis regulation
Mitosis
Chromatin histone methylation
Hormone−related
EGFR signaling
WNT signaling
p53 pathway
Chromatin histone acetylation
RTK signaling
ERK MAPK signaling
PI3K/AKT/mTOR signaling

(a)


Gene Set
GSEA NES
5
0
−5
Apoptosis regulation
EGFR signaling
stability and degradation
RTK signaling
Cell cycle
Genome integrity
ABL signaling
DNA replication
Mitosis
JNK and p38 signaling
PI3K/AKTOR signaling
WNT signaling
tatin histone methylation
p53 pathway
ERK MAPK signaling
tatin histone acetylation
IGF1R signaling
Metabolism
Cytoskeleton
Hormone−related

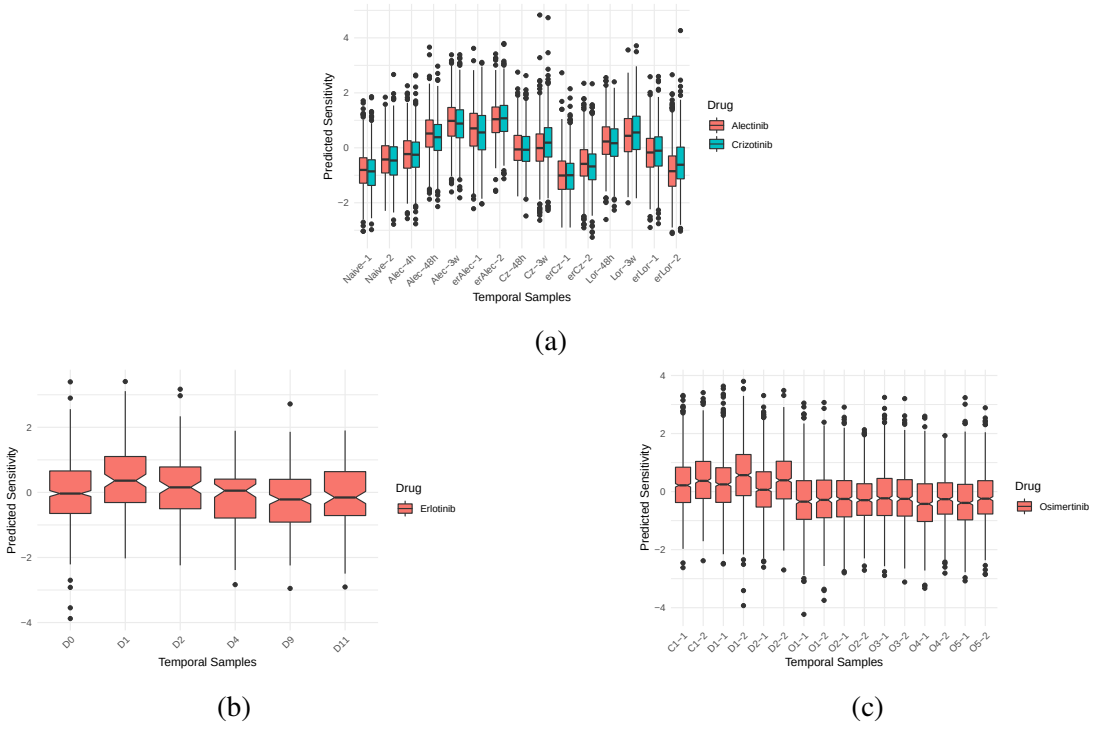
(b)

Figure D9



(a) (b)

Figure D10: **Navitoclax-Tanespimycin combination profiles showing synergistic activity using SynergyFinder in parental (a) and Gefitinib resistant cell-lines (b).**



(a)

(b) (c)

Figure D11: **scRNA-Seq temporal dataset quantifying transcriptional dynamics during resistance evolution**. (a) Alectinib, Crizotinib and Lorlatinib treatment over 6 months of *EML4-ALK*+ cell-lines showing overlap of Alectinib treatment and opposite predictions in Crizotinib treatet data. (b) Erlotinib treatment in *EGFR*+ cell-lines which overlaps with increasing resistance predictions. (c) Osimertinib treatment in *EGFR*+ cell-lines showing no association of drug sensitivity with sampling time

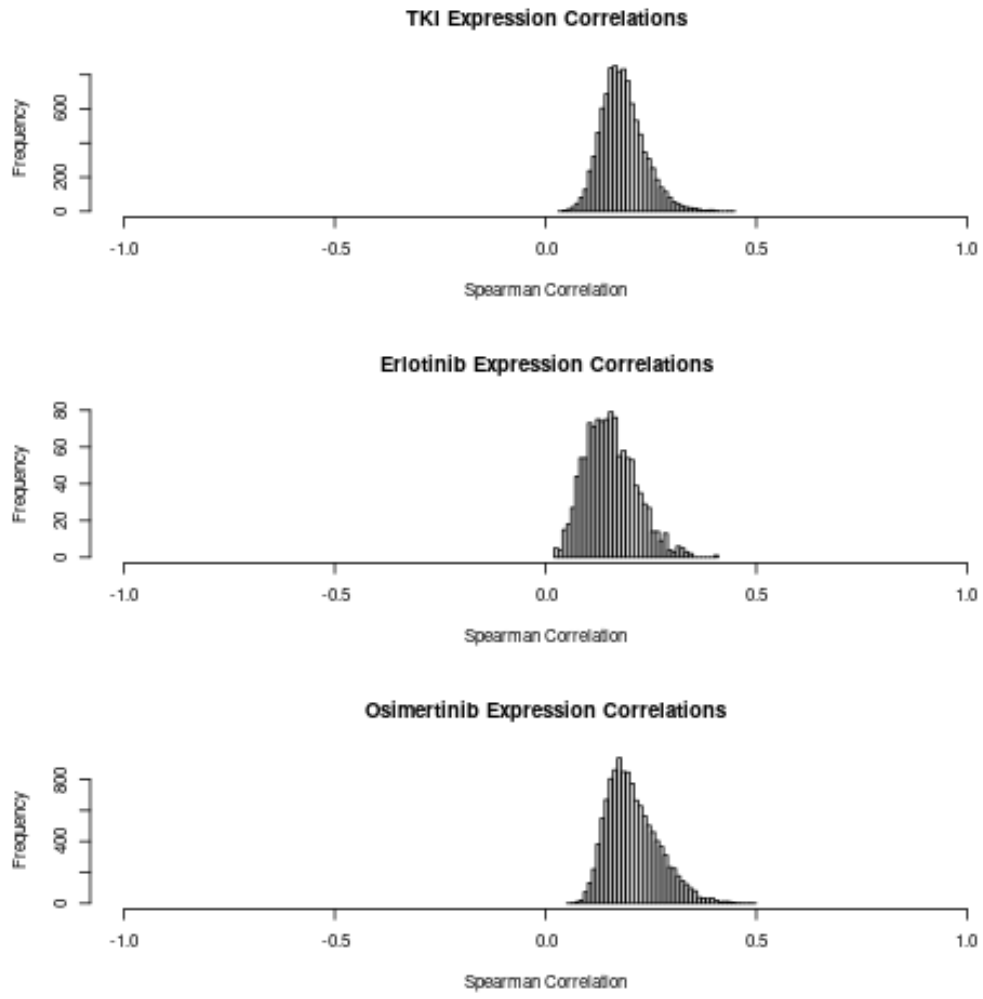


Figure D12: **Expression correlations across scRNA-Seq datasets**. Prediction comparisons showing reduced capacity of scRNA-Seq encoding.

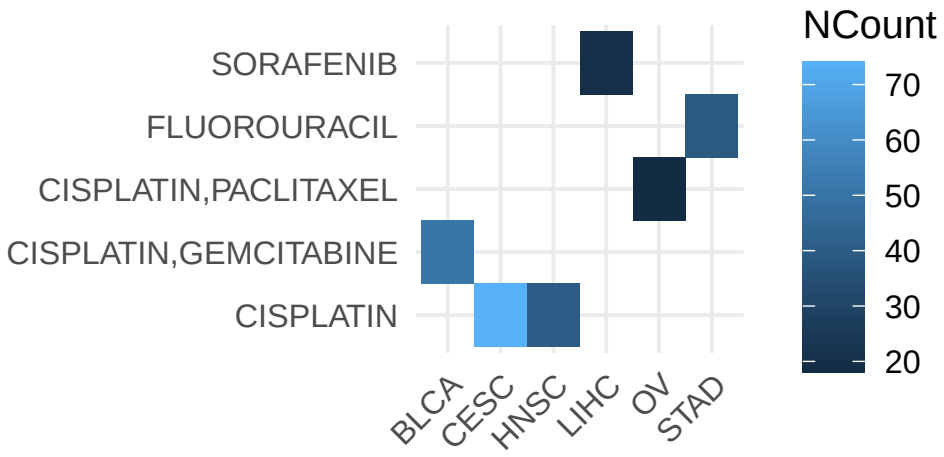


Figure D13: **Heatmap showing the cancer type-drug combination clinical data available for time-to-event modeling.** We have filtered out cancer type-drug combinations with < 5 events defined as progression. Colorbar represents the total number of patient observations.

BIBLIOGRAPHY

[1] Kenichi Yoshida et al. “Frequent pathway mutations of splicing machinery in myelodysplasia”. In: *Nature* 478.7367 (2011), pp. 64–69.

[2] Hideki Makishima et al. “Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis”. In: *Blood, The Journal of the American Society of Hematology* 119.14 (2012), pp. 3203–3210.

[3] Joseph D Khoury et al. “The 5th edition of the world health organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms”. In: *Leukemia* (2022), pp. 1–17.

[4] Hartmut Döhner et al. “Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel”. In: *Blood, The Journal of the American Society of Hematology* 129.4 (2017), pp. 424–447.

[5] Hartmut Döhner et al. “Diagnosis and Management of AML in Adults: 2022 ELN Recommendations from an International Expert Panel”. In: *Blood* (July 2022). blood.2022016867. ISSN: 0006-4971. DOI: 10.1182/blood.2022016867. eprint: https://ashpublications.org/blood/article-pdf/doi/10.1182/blood.2022016867/1906555/blood.2022016867.pdf. URL: https://doi.org/10.1182/blood.2022016867.

[6] Peter L Greenberg et al. “Revised international prognostic scoring system for myelodysplastic syndromes”. In: *Blood, The Journal of the American Society of Hematology* 120.12 (2012), pp. 2454–2465.

[7] Elsa Bernard et al. “Molecular International Prognostic Scoring System for Myelodysplastic Syndromes”. In: *NEJM Evidence* 1.7 (2022), EVIDoa2200008. DOI: 10.1056/EVIDoa2200008. eprint: https://evidence.nejm.org/doi/pdf/10.1056/EVIDoa2200008. URL: https://evidence.nejm.org/doi/abs/10.1056/EVIDoa2200008.

[8] Matteo Bersanelli et al. “Classification and personalized prognostic assessment on the basis of clinical and genomic features in myelodysplastic syndromes”. In: *Journal of Clinical Oncology* 39.11 (2021), pp. 1223–1233.

[9] Aziz Nazha et al. “Personalized prediction model to risk stratify patients with myelodysplastic syndromes”. In: *Journal of Clinical Oncology* 39.33 (2021), pp. 3737–3746.

[10] Elli Papaemmanuil et al. “Genomic classification and prognosis in acute myeloid leukemia”. In: *New England Journal of Medicine* 374.23 (2016), pp. 2209–2221.

[11] Alfonso Quintás-Cardama and Jorge Cortes. “Molecular biology of bcr-abl1–positive chronic myeloid leukemia”. In: *Blood, The Journal of the American Society of Hematology* 113.8 (2009), pp. 1619–1630.

[12] Andreas Hochhaus et al. “Long-term outcomes of imatinib treatment for chronic myeloid leukemia”. In: *New England Journal of Medicine* 376.10 (2017), pp. 917–927.

[13] Daniel A Arber et al. “The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia”. In: *Blood, The Journal of the American Society of Hematology* 127.20 (2016), pp. 2391–2405.

[14] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. “Intra-tumour heterogeneity: a looking glass for cancer?” In: *Nature reviews cancer* 12.5 (2012), pp. 323–334.

[15] Nicholas J Szerlip et al. “Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response”. In: *Proceedings of the National Academy of Sciences* 109.8 (2012), pp. 3041–3046.

[16] Charles Swanton et al. “Chromosomal instability determines taxane response”. In: *Proceedings of the National Academy of Sciences* 106.21 (2009), pp. 8671–8676.

[17] Ansuman T Satpathy et al. “Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion”. In: *Nature biotechnology* 37.8 (2019), pp. 925–936.

[18] Yongzheng Cong et al. “Ultrasensitive single-cell proteomics workflow identifies> 1000 protein groups per mammalian cell”. In: *Chemical Science* 12.3 (2021), pp. 1001–1006.

[19] Bogdan Budnik et al. “SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation”. In: *Genome biology* 19.1 (2018), pp. 1–12.

[20] Junyue Cao et al. “Joint profiling of chromatin accessibility and gene expression in thousands of single cells”. In: *Science* 361.6409 (2018), pp. 1380–1385.

[21] Iain C Macaulay et al. “G&T-seq: parallel sequencing of single-cell genomes and transcriptomes”. In: *Nature methods* 12.6 (2015), pp. 519–522.

[22] Vanessa M Peterson et al. “Multiplexed quantification of proteins and transcripts in single cells”. In: *Nature biotechnology* 35.10 (2017), pp. 936–939.

[23] Jong Kyoung Kim et al. “Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression”. In: *Nature communications* 6.1 (2015), pp. 1–9.

[24] Peng Qiu. “Embracing the dropouts in single-cell RNA-seq analysis”. In: *Nature communications* 11.1 (2020), pp. 1–9.

[25] Wei Vivian Li and Jingyi Jessica Li. “An accurate and robust imputation method scImpute for single-cell RNA-seq data”. In: *Nature communications* 9.1 (2018), pp. 1–9.

[26] Wuming Gong et al. “DrImpute: imputing dropout events in single cell RNA sequencing data”. In: *BMC bioinformatics* 19.1 (2018), pp. 1–10.

[27] Christoph Hafemeister and Rahul Satija. “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. In: *Genome Biology* 20.1 (2019), pp. 1–15.

[28] Rhonda Bacher et al. “SCnorm: robust normalization of single-cell RNA-seq data”. In: *Nature methods* 14.6 (2017), pp. 584–586.

[29] Aaron TL Lun, Karsten Bach, and John C Marioni. “Pooling across cells to normalize single-cell RNA sequencing data with many zero counts”. In: *Genome biology* 17.1 (2016), p. 75.

[30] Laurens van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[31] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[32] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[33] Daniel Nichol et al. “Antibiotic collateral sensitivity is contingent on the repeatability of evolution”. In: *Nature communications* 10.1 (2019), pp. 1–10.

[34] Jasmine Foo and Franziska Michor. “Evolution of resistance to targeted anti-cancer therapies during continuous and pulsed administration strategies”. In: *PLoS computational biology* 5.11 (2009), e1000557.

[35] Jingsong Zhang et al. “Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer”. In: *Nature communications* 8.1 (2017), pp. 1–9.

[36] Lejla Imamovic and Morten OA Sommer. “Use of collateral sensitivity networks to design drug cycling protocols that avoid resistance development”. In: *Science translational medicine* 5.204 (2013), 204ra132–204ra132.

[37] Andrew Dhawan et al. “Collateral sensitivity networks reveal evolutionary instability and novel treatment strategies in ALK mutated non-small cell lung cancer”. In: *Scientific reports* 7.1 (2017), pp. 1–9.

[38] Jessica A Scarborough et al. “Identifying States of Collateral Sensitivity during the Evolution of Therapeutic Resistance in Ewing’s Sarcoma”. In: *Iscience* 23.7 (2020), p. 101293.

[39] Jeff Maltas and Kevin B Wood. “Pervasive and diverse collateral sensitivity profiles inform optimal strategies to limit antibiotic resistance”. In: *PLoS biology* 17.10 (2019), e3000515.

[40] Sewall Wright et al. “The roles of mutation, inbreeding, crossbreeding, and selection in evolution”. In: (1932).

[41] Stuart Kauffman and Simon Levin. “Towards a general theory of adaptive walks on rugged landscapes”. In: *Journal of theoretical Biology* 128.1 (1987), pp. 11–45.

[42] Eshan S. King et al. “Fitness seascapes facilitate the prediction of therapy resistance under time-varying selection”. In: *bioRxiv* (2022). DOI: 10.1101/2022.06.10.495696. eprint: https://www.biorxiv.org/content/early/2022/06/12/2022.06.10.495696.full.pdf. URL: https://www.biorxiv.org/content/early/2022/06/12/2022.06.10.495696.

[43] Laura Cantini et al. “Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer”. In: *Nature communications* 12.1 (2021), pp. 1–12.

[44] Wanjuan Yang et al. “Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells”. In: *Nucleic acids research* 41.D1 (2012), pp. D955–D961.

[45] Aviad Tsherniak et al. “Defining a cancer dependency map”. In: *Cell* 170.3 (2017), pp. 564–576.

[46] Hartmut Döhner, Daniel J Weisdorf, and Clara D Bloomfield. “Acute myeloid leukemia”. In: *New England Journal of Medicine* 373.12 (2015), pp. 1136–1152.

[47] David Grimwade et al. “Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials”. In: *Blood, The Journal of the American Society of Hematology* 116.3 (2010), pp. 354–365.

[48] David Grimwade, Adam Ivey, and Brian JP Huntly. “Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance”. In: *Blood, The Journal of the American Society of Hematology* 127.1 (2016), pp. 29–41.

[49] Cancer Genome Atlas Research Network. “Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia”. In: *New England Journal of Medicine* 368.22 (2013), pp. 2059–2074.

[50] Jay P Patel et al. “Prognostic relevance of integrated genetic profiling in acute myeloid leukemia”. In: *New England Journal of Medicine* 366.12 (2012), pp. 1079–1089.

[51] John S Welch et al. “The origin and evolution of mutations in acute myeloid leukemia”. In: *Cell* 150.2 (2012), pp. 264–278.

[52] R Coleman Lindsley et al. “Acute myeloid leukemia ontogeny is defined by distinct somatic mutations”. In: *Blood, The Journal of the American Society of Hematology* 125.9 (2015), pp. 1367–1376.

[53] Matthew J Walter et al. “Clonal architecture of secondary acute myeloid leukemia”. In: *New England Journal of Medicine* 366.12 (2012), pp. 1090–1098.

[54] Brunangelo Falini et al. “Multilineage dysplasia has no impact on biologic, clinicopathologic, and prognostic features of AML with mutated nucleophosmin (NPM1)”. In: *Blood, The Journal of the American Society of Hematology* 115.18 (2010), pp. 3776–3786.

[55] Olga K Weinberg et al. “Association of mutations with morphological dysplasia in de novo acute myeloid leukemia without 2016 WHO Classification-defined cytogenetic abnormalities”. In: *haematologica* 103.4 (2018), p. 626.

[56] Jeffrey W Tyner et al. “Functional genomic landscape of acute myeloid leukaemia”. In: *Nature* 562.7728 (2018), pp. 526–531.

[57] Yasunobu Nagata et al. “Invariant patterns of clonal succession determine specific clinical features of myelodysplastic syndromes”. In: *Nature communications* 10.1 (2019), pp. 1–14.

[58] Cassandra M Hirsch et al. “Consequences of mutant TET2 on clonality and subclonal hierarchy”. In: *Leukemia* 32.8 (2018), pp. 1751–1761.

[59] Manja Meggendorfer et al. “Molecular analysis of myelodysplastic syndrome with isolated deletion of the long arm of chromosome 5 reveals a specific spectrum of molecular mutations with prognostic impact: a study on 123 patients and 27 genes”. In: *Haematologica* 102.9 (2017), p. 1502.

[60] Sabit Delic et al. “Application of an NGS-based 28-gene panel in myeloproliferative neoplasms reveals distinct mutation patterns in essential thrombocythaemia, primary myelofibrosis and polycythaemia vera”. In: *British journal of haematology* 175.3 (2016), pp. 419–426.

[61] A Kohlmann et al. “Monitoring of residual disease by next-generation deep-sequencing of RUNX1 mutations can identify acute myeloid leukemia patients with resistant disease”. In: *Leukemia* 28.1 (2014), pp. 129–137.

[62] Laura Palomo et al. “Molecular landscape and clonal architecture of adult myelodysplastic/myeloproliferative neoplasms”. In: *Blood* 136.16 (2020), pp. 1851–1862.

[63] Arthur White and Thomas Brendan Murphy. “BayesLCA: An R package for Bayesian latent class analysis”. In: *Journal of Statiscal Software* 61.13 (2014), pp. 1–28.

[64] Andrew Kuykendall et al. “Acute myeloid leukemia: the good, the bad, and the ugly”. In: *American Society of Clinical Oncology Educational Book* 38 (2018), pp. 555–573.

[65] Marina Dıaz-Beyá et al. “The prognostic value of multilineage dysplasia in de novo acute myeloid leukemia patients with intermediate-risk cytogenetics is dependent on NPM1 mutational status”. In: *Blood, The Journal of the American Society of Hematology* 116.26 (2010), pp. 6147–6148.

[66] Frank G Rücker et al. “TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome”. In: *Blood, The Journal of the American Society of Hematology* 119.9 (2012), pp. 2114–2121.

[67] Klaus H Metzeler et al. “Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia”. In: *Blood, The Journal of the American Society of Hematology* 128.5 (2016), pp. 686–698.

[68] Konstanze Döhner et al. “Impact of NPM1/FLT3-ITD genotypes defined by the 2017 European LeukemiaNet in patients with acute myeloid leukemia”. In: *Blood* 135.5 (2020), pp. 371–380.

[69] VI Gaidzik et al. “RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features”. In: *Leukemia* 30.11 (2016), pp. 2160–2168.

[70] Mario Cazzola, Matteo G Della Porta, and Luca Malcovati. “The genetic basis of myelodysplasia and its clinical relevance”. In: *Blood, The Journal of the American Society of Hematology* 122.25 (2013), pp. 4021–4034.

[71] Maria Teresa Voso and Carmelo Gurnari. “Have we reached a molecular era in myelodysplastic syndromes?” In: *Hematology* 2021.1 (2021), pp. 418–427.

[72] Mario Cazzola. “Myelodysplastic syndromes”. In: *New England Journal of Medicine* 383.14 (2020), pp. 1358–1374.

[73] John M Bennett. “Morphologic dysplasia in Myelodysplastic Syndromes: How accurate are morphologists?” In: *Leukemia research* 71 (2018), pp. 34–35.

[74] Matteo Giovanni Della Porta et al. “Minimal morphological criteria for defining bone marrow dysplasia: a basis for clinical implementation of WHO classification of myelodysplastic syndromes”. In: *Leukemia* 29.1 (2015), pp. 66–75.

[75] Kiran Naqvi et al. “Implications of discrepancy in morphologic diagnosis of myelodysplastic syndrome between referral and tertiary care centers”. In: *Blood, The Journal of the American Society of Hematology* 118.17 (2011), pp. 4690–4693.

[76] Xueyan Chen et al. “Comparison of myeloid blast counts and variant allele frequencies of gene mutations in myelodysplastic syndrome with excess blasts and secondary acute myeloid leukemia”. In: *Leukemia & Lymphoma* 62.5 (2021), pp. 1226–1233.

[77] T Haferlach et al. “Landscape of genetic lesions in 944 patients with myelodysplastic syndromes”. In: *Leukemia* 28.2 (2014), pp. 241–247.

[78] Hideki Makishima et al. “Dynamics of clonal evolution in myelodysplastic syndromes”. In: *Nature genetics* 49.2 (2017), pp. 204–212.

[79] Hussein Awada et al. “Personalized Risk Schemes and Machine Learning to Empower Genomic Prognostication Models in Myelodysplastic Syndromes”. In: *International Journal of Molecular Sciences* 23.5 (2022), p. 2802.

[80] Yasunobu Nagata et al. “Machine learning demonstrates that somatic mutations imprint invariant morphologic features in myelodysplastic syndromes”. In: *Blood* 136.20 (2020), pp. 2249–2262.

[81] Nathan Radakovich, Matthew Nagy, and Aziz Nazha. “Machine learning in haematological malignancies”. In: *The Lancet Haematology* 7.7 (2020), e541–e550.

[82] Hao Zhang et al. “Deep autoencoding topic model with scalable hybrid Bayesian inference”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.12 (2020), pp. 4306–4322.

[83] Ashwath Radhachandran et al. “A machine learning approach to predicting risk of myelodysplastic syndrome”. In: *Leukemia Research* 109 (2021), p. 106639.

[84] Yang Liang et al. “SRSF2 mutations drive oncogenesis by activating a global program of aberrant alternative splicing in hematopoietic cells”. In: *Leukemia* 32.12 (2018), pp. 2659–2671.

[85] Brian Reilly et al. “DNA methylation identifies genetically and prognostically distinct subtypes of myelodysplastic syndromes”. In: *Blood advances* 3.19 (2019), pp. 2845–2858.

[86] Kunihiko Hinohara and Kornelia Polyak. “Intratumoral Heterogeneity: More Than Just Mutations”. In: *Trends in cell biology* (2019).

[87] Antonija Kreso and John E Dick. “Evolution of the cancer stem cell model”. In: *Cell stem cell* 14.3 (2014), pp. 275–291.

[88] Rebecca A Burrell and Charles Swanton. “Tumour heterogeneity and the evolution of polyclonal drug resistance”. In: *Molecular oncology* 8.6 (2014), pp. 1095–1111.

[89] Daniela F Quail and Johanna A Joyce. “Microenvironmental regulation of tumor progression and metastasis”. In: *Nature medicine* 19.11 (2013), p. 1423.

[90] Artem Kaznatcheev et al. “Fibroblasts and alectinib switch the evolutionary games played by non-small cell lung cancer”. In: *Nature ecology & evolution* 3.3 (2019), pp. 450–456.

[91] Mei-Chong Wendy Lee et al. “Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing”. In: *Proceedings of the National Academy of Sciences* 111.44 (2014), E4726–E4735.

[92] Kyu-Tae Kim et al. “Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells”. In: *Genome biology* 16.1 (2015), p. 127.

[93] Itay Tirosh et al. “Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma”. In: *Nature* 539.7628 (2016), p. 309.

[94] Ankur Sharma et al. “Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy”. In: *Nature communications* 9.1 (2018), p. 4931.

[95] Sung Pil Hong et al. “Single-cell Transcriptomics reveals multi-step adaptations to endocrine therapy”. In: *bioRxiv* (2018), p. 485136.

[96] Kyle J Card et al. “Historical contingency in the evolution of antibiotic resistance after decades of relaxed selection”. In: *PLoS biology* 17.10 (2019), e3000397.

[97] Bo Lv et al. “Single-cell RNA sequencing reveals regulatory mechanism for trophoblast cell-fate divergence in human peri-implantation conceptuses”. In: *PLoS biology* 17.10 (2019), e3000187.

[98] Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. In: *bioRxiv* (2020).

[99] Tim Stuart et al. “Comprehensive integration of single-cell data”. In: *Cell* 177.7 (2019), pp. 1888–1902.

[100] Andrew Butler et al. “Integrating single-cell transcriptomic data across different conditions, technologies, and species”. In: *Nature biotechnology* 36.5 (2018), pp. 411–420.

[101] Rahul Satija et al. “Spatial reconstruction of single-cell gene expression data”. In: *Nature biotechnology* 33.5 (2015), pp. 495–502.

[102] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome biology* 19.1 (2018), pp. 1–5.

[103] Páll Melsted, Vasilis Ntranos, and Lior Pachter. “The barcode, UMI, set format and BUStools”. In: *Bioinformatics* 35.21 (2019), pp. 4472–4473.

[104] Páll Melsted et al. “Modular and efficient pre-processing of single-cell RNA-seq”. In: *BioRxiv* (2019), p. 673285.

[105] Wouter Saelens et al. “A comparison of single-cell trajectory inference methods”. In: *Nature biotechnology* 37.5 (2019), pp. 547–554.

[106] Luyi Tian et al. “Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments”. In: *Nature methods* 16.6 (2019), pp. 479–487.

[107] Nir Yosef and Aviv Regev. “Impulse control: temporal dynamics in gene transcription”. In: *Cell* 144.6 (2011), pp. 886–896.

[108] Arno Steinacher et al. “Nonlinear dynamics in gene regulation promote robustness and evolvability of gene expression levels”. In: *PloS one* 11.4 (2016), e0153295.

[109] Michael J Lee et al. “Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks”. In: *Cell* 149.4 (2012), pp. 780–794.

[110] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[111] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[112] Martin Palazzo, Pierre Beauseroy, and Patricio Yankilevich. “A pan-cancer somatic mutation embedding using autoencoders”. In: *BMC bioinformatics* 20.1 (2019), pp. 1–10.

[113] Ze Xiao and Yue Deng. “Graph embedding-based novel protein interaction prediction via higher-order graph convolutional network”. In: *PloS one* 15.9 (2020), e0238915.

[114] Jiarui Ding and Aviv Regev. “Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces”. In: *BioRxiv* (2019), p. 853457.

[115] Gökcen Eraslan et al. “Single-cell RNA-seq denoising using a deep count autoencoder”. In: *Nature communications* 10.1 (2019), pp. 1–14.

[116] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific reports* 9.1 (2019), pp. 1–12.

[117] F Alexander Wolf et al. “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells”. In: *Genome biology* 20.1 (2019), pp. 1–9.

[118] Sabrina Rashid et al. “Dhaka: variational autoencoder for unmasking tumor heterogeneity from single cell genomic data”. In: *bioRxiv* (2018), p. 183863.

[119] Laleh Haghverdi, Florian Buettner, and Fabian J Theis. “Diffusion maps for high-dimensional single-cell analysis of differentiation data”. In: *Bioinformatics* 31.18 (2015), pp. 2989–2998.

[120] Gregory W Schwartz et al. “TooManyCells identifies and visualizes relationships of single-cell clades”. In: *Nature methods* 17.4 (2020), pp. 405–413.

[121] Kelly Street et al. “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics”. In: *BMC genomics* 19.1 (2018), p. 477.

[122] Manu Setty et al. “Characterization of cell fate probabilities in single-cell data with Palantir”. In: *Nature biotechnology* 37.4 (2019), pp. 451–460.

[123] Xiaojie Qiu et al. “Reversed graph embedding resolves complex single-cell trajectories”. In: *Nature methods* 14.10 (2017), p. 979.

[124] Geoffrey Schiebinger et al. “Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming”. In: *Cell* 176.4 (2019), pp. 928–943.

[125] Robert Vander Velde et al. “Resistance to targeted therapies as a multifactorial, gradual adaptation to inhibitor specific selective pressures”. In: *Nature communications* 11.1 (2020), pp. 1–13.

[126] Wei-Lin Qiu et al. “Deciphering pancreatic islet β cell and α cell maturation pathways and characteristic features at the single-cell level”. In: *Cell metabolism* 25.5 (2017), pp. 1194–1205.

[127] Detu Zhu et al. “Single-cell transcriptome analysis reveals estrogen signaling coordinately augments one-carbon, polyamine, and purine synthesis in breast cancer”. In: *Cell reports* 25.8 (2018), pp. 2285–2298.

[128] Hansruedi Mathys et al. “Temporal tracking of microglia activation in neurodegeneration at single-cell resolution”. In: *Cell reports* 21.2 (2017), pp. 366–380.

[129] Haotian Zhuang, Huimin Wang, and Zhicheng Ji. “findPC: An R package to automatically select the number of principal components in single-cell analysis”. In: *Bioinformatics* 38.10 (2022), pp. 2949–2951.

[130] Mo Huang et al. “SAVER: gene expression recovery for single-cell RNA sequencing”. In: *Nature methods* 15.7 (2018), pp. 539–542.

[131] David Van Dijk et al. “Recovering gene interactions from single-cell data using data diffusion”. In: *Cell* 174.3 (2018), pp. 716–729.

[132] Wenpin Hou et al. “A systematic evaluation of single-cell RNA-sequencing imputation methods”. In: *Genome biology* 21.1 (2020), pp. 1–30.

[133] Cody N Heiser and Ken S Lau. “A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques”. In: *Cell reports* 31.5 (2020), p. 107576.

[134] Chuner Guo et al. “CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics”. In: *Genome biology* 20.1 (2019), p. 90.

[135] Wenjun Kong et al. “CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution”. In: *Nature protocols* 15.3 (2020), pp. 750–772.

[136] Patricia Jaaks et al. “Effective drug combinations in breast, colon and pancreatic cancer cells”. In: *Nature* (2022), pp. 1–8.

[137] Susan L Holbeck et al. “The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity”. In: *Cancer research* 77.13 (2017), pp. 3564–3576.

[138] Tianqi Chen et al. “Xgboost: extreme gradient boosting”. In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[139] Pavel Sidorov et al. “Predicting synergism of cancer drug combinations using NCI-ALMANAC data”. In: *Frontiers in chemistry* (2019), p. 509.

[140] Peiran Jiang et al. “Deep graph embedding for prioritizing synergistic anticancer drug combinations”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 427–438.

[141] Jessica Scarborough, Andrew Dhawan, and Jacob Scott. *Exploiting convergent evolution to derive a cisplatin sensitivity gene expression signature in epithelial based cancer*. 2020.

[142] Patricia M Schnepp et al. “Transcription factor network analysis based on single cell RNA-seq identifies that Trichostatin-a reverses docetaxel resistance in prostate Cancer”. In: *BMC cancer* 21.1 (2021), pp. 1–14.

[143] Lourdes Hontecillas-Prieto et al. “Synergistic enhancement of cancer therapy using HDAC inhibitors: opportunity for clinical trials”. In: *Frontiers in genetics* 11 (2020), p. 578011.

[144] Amila Suraweera, Kenneth J O’Byrne, and Derek J Richard. “Combination therapy with histone deacetylase inhibitors (HDACi) for the treatment of cancer: achieving the full therapeutic potential of HDACi”. In: *Frontiers in oncology* 8 (2018), p. 92.

[145] Ayana Sawai et al. “Inhibition of Hsp90 down-regulates mutant epidermal growth factor receptor (EGFR) expression and sensitizes EGFR mutant tumors to paclitaxel”. In: *Cancer research* 68.2 (2008), pp. 589–596.

[146] Sho Watanabe et al. “HSP90 inhibition overcomes EGFR amplification-induced resistance to third-generation EGFR-TKIs”. In: *Thoracic cancer* 12.5 (2021), pp. 631–642.

[147] Panagiotis K Karkoulis et al. “Targeted inhibition of heat shock protein 90 disrupts multiple oncogenic signaling pathways, thus inducing cell cycle arrest and programmed cell death in human urinary bladder cancer cell lines”. In: *Cancer cell international* 13.1 (2013), pp. 1–16.

[148] Yixuan Gong et al. “Induction of BIM is essential for apoptosis triggered by EGFR kinase inhibitors in mutant EGFR-dependent lung adenocarcinomas”. In: *PLoS medicine* 4.10 (2007), e294.

[149] Mark S Cragg et al. “Gefitinib-induced killing of NSCLC cell lines expressing mutant EGFR requires BIM and can be enhanced by BH3 mimetics”. In: *PLoS medicine* 4.10 (2007), e316.

[150] Jianfang Liu et al. “An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics”. In: *Cell* 173.2 (2018), pp. 400–416.

[151] Yukie Kashima et al. “Single-cell analyses reveal diverse mechanisms of resistance to EGFR tyrosine kinase inhibitors in lung cancer”. In: *Cancer research* 81.18 (2021), pp. 4835–4848.

[152] Nadia Godin-Heymann et al. “The T790M “gatekeeper” mutation in EGFR mediates resistance to low concentrations of an irreversible EGFR inhibitor”. In: *Molecular cancer therapeutics* 7.4 (2008), pp. 874–879.

[153] Ricard Argelaguet et al. “MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data”. In: *Genome biology* 21.1 (2020), pp. 1–17.

[154] Mingbo Cheng, Zhijian Li, and Ivan Gesteira Costa Filho. “MOJITOO: a fast and universal method for integration of multimodal single cell data”. In: *bioRxiv* (2022).

[155] Deirdra Venney, Adone Mohd-Sarip, and Ken I Mills. “The impact of epigenetic modifications in myeloid malignancies”. In: *International Journal of Molecular Sciences* 22.9 (2021), p. 5013.

[156] Ming-En Lin et al. “Dynamics of DNMT3A mutation and prognostic relevance in patients with primary myelodysplastic syndrome”. In: *Clinical epigenetics* 10.1 (2018), pp. 1–12.

[157] Jinming Song et al. “Comparison of SF3B1/DNMT3A Comutations with DNMT3A or SF3B1 mutation alone in Myelodysplastic syndrome and clonal Cytopenia of undetermined significance”. In: *American journal of clinical pathology* 154.1 (2020), pp. 48–56.

[158] Waled Bahaj et al. “The Drive to Acquire Biallelic Hits Inversely Correlates with the Functional Impact of the Primary TP53 Lesion: The Complexity of TP53 Role Assessment”. In: *Blood* 138 (2021), p. 3322.

[159] Elsa Bernard et al. “Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes”. In: *Nature medicine* 26.10 (2020), pp. 1549–1556.

[160] Tim Grob et al. “Molecular characterization of mutant TP53 acute myeloid leukemia and high-risk myelodysplastic syndrome”. In: *Blood, The Journal of the American Society of Hematology* 139.15 (2022), pp. 2347–2354.

[161] Borahm Kim et al. “Somatic mosaic truncating mutations of PPM1D in blood can result from expansion of a mutant clone under selective pressure of chemotherapy”. In: *PloS one* 14.6 (2019), e0217521.

[162] Thomas Kindler. “CHIPing out PPM1D-mutant hematopoiesis”. In: *Blood, The Journal of the American Society of Hematology* 132.11 (2018), pp. 1087–1088.

[163] Abhay Singh et al. “Mutant PPM1D-and TP53-Driven hematopoiesis populates the hematopoietic compartment in response to peptide receptor radionuclide therapy”. In: *JCO Precision Oncology* 6 (2022), e2100309.

[164] Juan Ramon Gonzalez Garcia and Juan Pablo Meza-Espinoza. “Use of the international system for human cytogenetic nomenclature (ISCN)”. In: *Blood* 108.12 (2006), pp. 3952–3953.

[165] Thomas Liehr. “International system for human cytogenetic or cytogenomic nomenclature (ISCN): Some thoughts”. In: *Cytogenetic and Genome Research* 161.5 (2021), pp. 223–224.

[166] Pierre Baldi. “Autoencoders, unsupervised learning, and deep architectures”. In: *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings. 2012, pp. 37–49.

[167] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.

[168] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[169] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[170] Ronald R Coifman et al. “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps”. In: *Proceedings of the national academy of sciences* 102.21 (2005), pp. 7426–7431.

[171] Philipp Angerer et al. “destiny: diffusion maps for large-scale single-cell data in R”. In: *Bioinformatics* 32.8 (2016), pp. 1241–1243.

[172] Qi Mao et al. “Dimensionality reduction via graph structure learning”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 765–774.

[173] Qi Mao et al. “SimplePPT: A simple principal tree algorithm”. In: *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM. 2015, pp. 792–800.