Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Μηχανολόγων Μηχανικών
Τομέας Μηχανολογικών Κατασκευών & Αυτομάτου Ελέγχου

Development of analytical frameworks for the integration of molecular data views to understand the biology of myeloid malignancies

Διδακτορική Διατριβή

Γεώργιος Ασιμομήτης

ΔΙΠΛΩΜΑΤΟΥΧΟΥ ΗΛΕΚΤΡΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ ΕΜΠ

ΕΠΙΒΛΕΠΩΝ:
Λεωνίδας Αλεξόπουλος, Καθηγητής, ΕΜΠ

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Μηχανολόγων Μηχανικών
Τομέας Μηχανολογικών Κατασκευών & Αυτομάτου Ελέγχου

Development of analytical frameworks for the integration of molecular data views to understand the biology of myeloid malignancies

Διδακτορική Διατριβή

Γεώργιος Ασιμομήτης

ΔΙΠΛΩΜΑΤΟΥΧΟΥ ΗΛΕΚΤΡΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ ΕΜΠ

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Λ. Αλεξόπουλος, Καθ. ΕΜΠ (Επιβλέπων)

Ε. Παπαεμμανουήλ, Αν. Καθ. MSKCC

Ι. Κοτσιανίδης, Καθ. ΔΠΘ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Λ. Αλεξόπουλος, Καθ. ΕΜΠ (Επιβλέπων)

Ε. Παπαεμμανουήλ, Αν. Καθ. MSKCC

Ι. Κοτσιανίδης, Καθ. ΔΠΘ

Χ. Μανόπουλος, Επ. Καθ. ΕΜΠ

Ε. Παπαπέτρου, Καθ. Mt Sinai

Α. Σπυριδωνίδης, Καθ. Παν. Πατρών

Α. Τσιρίγος, Καθ. NYU

Athens, January 2024

Η έγκριση της διδακτορικής διατριβής από την Ανώτατη Σχολή Μηχανολόγων Μηχανικών του Ε.Μ. Πολυτεχνείου δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα (Ν. 5343/1932, Άρθρο 202)

Prologue

The research for this PhD dissertation was carried out under the joint supervision of the Professor of Mechanical Engineering at the National Technical University of Athens, Leonidas G. Alexopoulos and the Professor of Computational Oncology at the Memorial Sloan Kettering Cancer Center in New York, Elli Papaemmanuil. My research has been funded by the MDS Foundation, Geoffrey Beene Foundation and MSK MIND.

I would like to sincerely thank both of my supervisors for their mentorship, guidance and advice throughout this journey. Working next to them has been a unique privilege and has equipped me with invaluable assets and experience for the continuation of my career. Feeling their faith and trust in my work allowed me to take initiative, develop my interests and skills as well as expand my creativity. Witnessing their way of thinking, acting, approaching daily matters and combining professional with personal life, has been a constant motivation for me to grow and advance not only scientifically but most importantly as an individual. Additionally, during these years I am grateful to have been surrounded by an exceptional group of colleagues and collaborators. First, I would like to thank Prof. Eirini Papapetrou. Our interaction has been a true lesson for me on how to perform high-quality research thoroughly and efficiently. Then, I would like to thank Drs. Andre Deslauriers and Maria Sirenko for generating the datasets I worked with in the context of my thesis. Additionally, I am thankful to Dr. Christos Fotis for his feedback on my work and computational advice. I would also like to express my gratitude to my labmates who comprise a really pleasant, easy-going and constructive environment to work in. Lastly, I am really appreciative of the support I gained from my partner, friends and family and especially my sister, who has been standing next to me every step of the way.

Georgios Asimomitis

Athens, January 2024

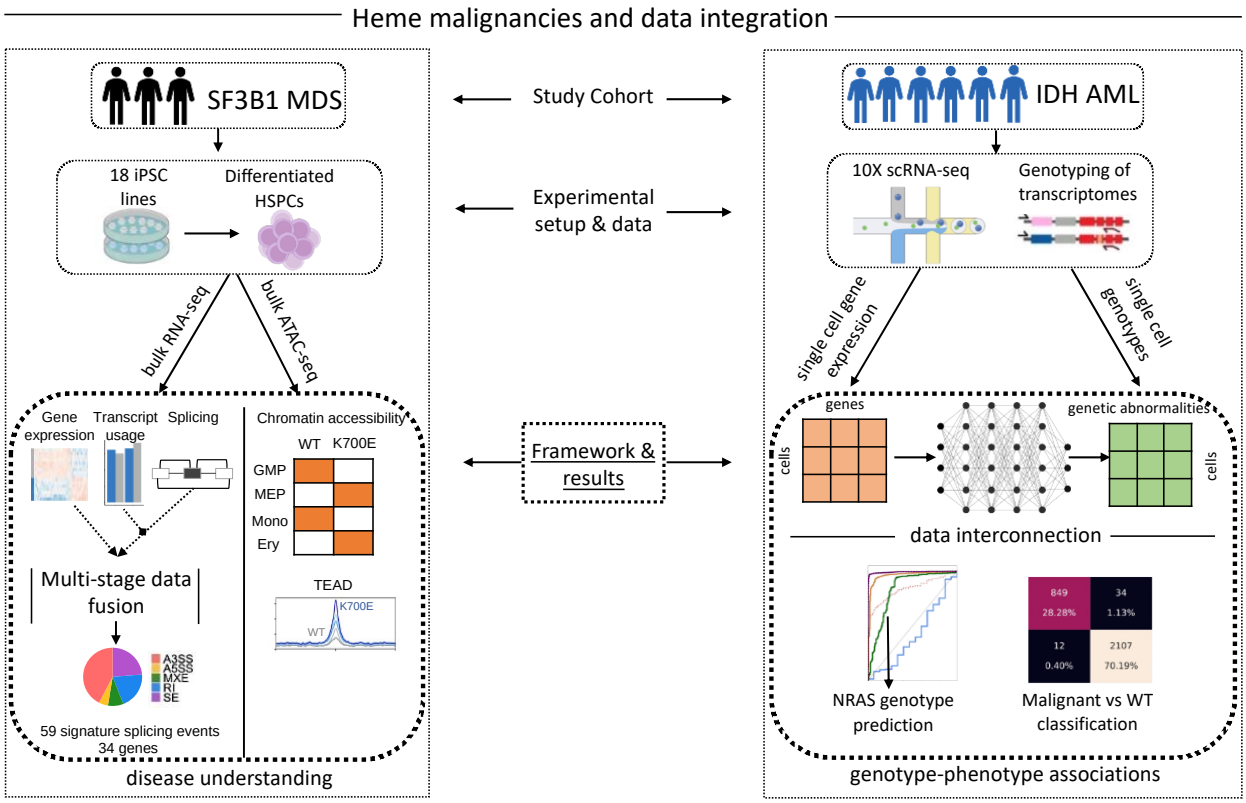
Summary

Myeloid malignancies consist of a heterogeneous spectrum of clonal stem cell disorders driven by genetic alterations, resulting in dysregulated hematopoiesis. The investigation of the mechanisms underpinning myeloid neoplasia relies primarily on experimental models of disease biology and the phenotyping of primary patient samples using emerging genomic technologies. In recognition of the increasing complexity, scale and dimensionality of the datasets generated by these approaches, this thesis focuses on the development of analytical frameworks that operate within and across different omics modalities (genomics, transcriptomics, epigenomics) and sequencing techniques (bulk and single cell), and set out to enhance the understanding of the underlying biology of myeloid neoplasms. Specifically, this work deploys principles from multi-view data fusion and interconnection to analyze signals in an integrative manner, aiming to elucidate molecular landscapes and assist the study of phenotypes at a genetic level.

Chapter 2 investigates the transcriptional repertoire and chromatin profile of *SF3B1*-mutated Myelodysplastic Syndromes (MDS), leveraging bulk RNA and ATAC sequencing data from patient-derived genetically matched normal and *SF3B1*-mutated induced pluripotent stem cell (iPSC) lines. We introduce a multi-stage fusion framework that merges signals from diverse data layers obtained from transcriptome sequencing (splicing, transcript usage, gene expression). The analytical framework developed as part of this work leads to the derivation of a splicing signature linked to 34 genes, which associates with the *SF3B1* mutational status of primary MDS patient cells. Additional unimodal chromatin accessibility analysis showed increased priming of *SF3B1* hematopoietic progenitors toward the megakaryocyte-erythroid lineage, as well as the enrichment of motifs from the TEA (TEAD) domain in accessible regions linked to genes with upregulated expression. Overall, chapter 2, applies a multi-stage fusion approach on transcriptomic data views to prioritize mis-spliced gene targets, and concurrently provides a formal overview of the *SF3B1*-mutated chromatin landscape and nominates transcriptional programs with putative roles in MDS disease biology.

Chapter 3 examines if single cell gene expression signals together with the computational capacity of neural networks are able to predict a cell's malignant status and subsequently its genotype for specific abnormalities in *IDH1/2*-mutated Acute Myeloid Leukemia (AML). To this end, using single cell RNA sequencing data from 50,026 cells, a feedforward neural network was trained to predict the cell's malignant or wild-type (WT) status in a binary fashion, achieving an accuracy of 98% on the holdout test set. Furthermore, in a multi-label setting, this work deploys a similar architecture to predict the mutational status of specific genomic abnormalities at the single cell level, showing a macro-average AUC ROC=0.84 and *NRAS* mutational status prediction AUC ROC=0.83 on the holdout test-set. Altogether, chapter 3 applies deep learning in a supervised context to explore the connection between single cell gene expression profiles and genotypes in *IDH1/2* AML and shows the potential of such modeling approaches in capturing meaningful genotype-phenotype relationships.

Graphical abstract



Extended summary

Introduction

Myeloid Neoplasms (MNs) constitute a continuum of clonal proliferative disorders, which are comprised of chronic phases including Myeloproliferative Neoplasms (MPN), Myelodysplastic Syndromes (MDS), and acute stages i.e. Acute Myeloid Leukemia (AML). MNs typically arise from the acquisition of genetic abnormalities that disrupt normal hematopoiesis. In recent years, genome profiling studies have delivered a detailed catalog of the somatic mutations in MNs. However, understanding the mechanisms leading to myeloid transformation and the effectors of disease biology relies on the development of experimental models (murine, cell based, organoids) as well as the phenotyping of primary patient samples. Such approaches are typically combined with profiling assays that analyze a sample's genome, transcriptome and epigenome. Next to the well-established bulk sequencing techniques, the more recent advancements of single cell technologies have also added to the routine yield of high-throughput and extensive omics datasets. These datasets contain distinct data views (representations or sets of features derived from the measured biomolecules either within or across modalities) that permit the investigation of molecular properties at multiple omic levels (genomic, transcriptomic or epigenomic). However, analyses focusing on a single data view do not lead to the full characterization of molecular landscapes and the establishment of genotype-phenotype associations. Thus, the development of analytical frameworks that allow for the integration and interpretation of multiple data views offers an opportunity to study the representation of different molecular layers and the relationships between them. In recognition of the increasing complexity, scale and dimensionality of the generated data as well as the need for gaining multi-faceted insights on disease behavior, this thesis sets out to develop analytical frameworks aiming to enhance the understanding of the underlying biology of myeloid neoplasms. Specifically, this work uses principles from multi-view data fusion and interconnection as a means of integrating signals within and between omics modalities (genomics, transcriptomics, epigenomics) either from bulk or single cell sequencing techniques. The presented analyses and the frameworks developed in the context of it, set out to elucidate molecular landscapes and assist the study of phenotypes at a genetic level, focusing on *SF3B1*-mutated MDS and *IDH1/2*-mutated AML correspondingly.

Patient-specific MDS-RS iPSCs define the mis-spliced transcript repertoire and chromatin landscape of *SF3B1*-mutant HSPCs

Background

Myelodysplastic syndromes (MDS) are myeloid malignancies characterized by ineffective hematopoiesis, blood cytopenias, and an increased risk of progression to AML. Recent sequencing studies have emphasized the role of mutations in splicing factor genes (*SF3B1*, *SRSF2*, *ZRSR2*, *U2AF1*) as initiating and MDS defining. Among these splicing factor genes, *SF3B1* is the most frequently mutated one in MDS (~ 24 % patients) and

defines a distinct nosologic entity, termed MDS with ring sideroblasts (MDS-RS). Mutations in *SF3B1* are commonly found as isolated events, mainly target the K700 hotspot and are associated with favorable outcomes. Despite the characterization of the molecular landscape in MDS, how such mutations drive disease pathogenesis and how they can inform clinical management remains unclear. In this study, by leveraging data from an experimental iPSC model, we explore the downstream consequences of the *SF3B1*^{K700E} mutation and its role in disease pathogenesis through the integration of multiple views from the transcriptome and the examination of the *SF3B1*^{K700E} chromatin accessibility landscape.

Data & Methods

Hereby, we used a panel of 18 genetically matched *SF3B1*^{K700E} and *SF3B1*^{WT} induced pluripotent stem cell (iPSC) lines derived from 3 MDS-RS patients who harbored isolated *SF3B1*^{K700E} mutations. For these iPSC lines, directed hematopoietic differentiation was performed using protocols from the Papapetrou laboratory and CD34^{+}/CD45^{+} human stem and progenitor cells (HSPCs) were collected for RNA and ATAC-sequencing. (We note that the generation of the data is not part of the current thesis). HSPC samples from 16 iPSC lines were included in the RNA-seq analyses after quality control of the raw data. RNA-seq reads were aligned and used for the quantification of transcript abundance and the generation of the gene counts. Differential gene expression analysis, transcript usage analysis and splicing analysis were conducted between *SF3B1*^{K700E} vs *SF3B1*^{WT} cells. To assess the impact of the *SF3B1*^{K700E} mutation at the exon, transcript and gene level, we combined signals from these 3 analyses in a multi-stage fusion setting. First, we identified the set of transcripts that contain the exons present in each differential splicing event. Then, we paired each differential splicing event with the set of differentially used transcripts. The pairs that belonged to genes with a statistically significant expression $\log_2$fc and contained a differential splicing event with an FDR value within the 20 lowermost ones, were considered as the "tier 1" set. From this set, we derived the mutant SF3B1 signature events and genes. Additionally, HSPC samples from 15 iPSC lines were included in the ATAC-seq analysis. After read alignment and quality control, we identified chromatin accessibility peaks and created an ATAC-seq atlas. This atlas was used for downstream differential accessibility analysis, correlation with the accessibility landscape of the normal hematopoietic hierarchy as well as motif enrichment analysis.

Results

Principal component analysis (PCA) and hierarchical clustering based on gene expression grouped the iPSC lines primarily by genotype (*SF3B1*^{K700E} vs *SF3B1*^{WT}). Additionally, differential analyses revealed 2737 differentially expressed genes, 1086 differentially used transcripts and 1829 differentially spliced events between *SF3B1*^{K700E} and *SF3B1*^{WT} cells. Integrating the signals from these analyses using our multi-stage fusion approach resulted in the derivation of a splicing signature consisting of 59 splicing events linked to 34 genes. We tested this signature against a published dataset of primary MDS patient samples (Pellagatti et al. Blood 2018). Specifically, PCA based on the inclusion level of the splicing events of our signature separated *SF3B1*-mutated MDS patients from patients without splicing factor mutations (SF-WT) or healthy individuals. Importantly, it identified one patient erroneously annotated as SF-WT that clustered together with the *SF3B1*-mutated patients. This patient had a previously overlooked 6 base pair (bp) in-frame deletion spanning the K700E hotspot. Comparing our ATAC-seq peak atlas to the chromatin accessibility profiles of primary

human cell types along the hematopoietic hierarchy (Corces et al. Nat Genetics 2016), we found that the chromatin landscape of *SF3B1*^{K700E} HSPCs resembled more that of megakaryocyte-erythroid progenitor cells (MEPs) and erythroid cells. Furthermore, motifs enriched in ATAC-Seq peaks more accessible in *SF3B1*^{K700E} cells that were linked to genes upregulated in *SF3B1*^{K700E} cells, included motifs of the TEAD transcription factor family. *TEAD2* and *TEAD4* were upregulated in *SF3B1*-mutant, compared to the WT iPSC-HSPCs and TEAD transcriptional activity, measured with a luciferase reporter construct, was higher in *SF3B1*^{K700E}, compared to *SF3B1*^{WT} iPSC-HSPCs. We did not find expression or activation of YAP or TAZ, which bind to DNA as a complex with TEAD upon Hippo pathway activation, suggesting a Hippo-independent increase of TEAD expression and activity in *SF3B1*^{K700E}.

Discussion

Powered by a data integration framework, this study assesses the combination of the effects of the *SF3B1*^{K700E} mutation across parallel levels of deregulation of the transcriptome towards deriving a tier-based classification of splicing events. Specifically, this framework systematically evaluates the relationships between the *SF3B1* mutation, differential splicing, transcript usage and gene expression and leads to a fully characterized *SF3B1*^{K700E} splicing signature. This signature includes several known gene candidates and is also able to identify atypical mutations involving the K700 hotspot. Furthermore, this study, shows, at the chromatin level, a potential "priming" of *SF3B1*^{K700E} HSPCs toward the erythroid over the myeloid lineage - a finding that may be related to the preferential involvement of the erythroid lineage in MDS and, in particular, MDS-RS. Lastly, our chromatin accessibility analyses lend support to a putative role for the TEAD TFs in the context of *SF3B1*^{K700E} mutation, a signal which warrants validation in future studies.

Predicting single cell genotypes from single cell expression profiles in AML using deep learning

Background

Approximately 30% of MDS patients eventually progress to AML, an aggressive blood cancer associated with rapid disease progression, poor response to therapy and dismal outcomes. AML is a genetically heterogeneous disease defined by the gradual accumulation of mutations. These are often characterized by specific gene by gene interactions, indicative of functional cooperativity, and result in genetically and clonally heterogeneous populations. This imposes a significant challenge in treating and ultimately curing the disease. Elucidating the role and effect of this diversity at the cellular phenotypes and disease biology requires: 1. molecular representations at the cellular level, which cannot be achieved by bulk sequencing approaches in primary tumor samples, and 2. analytical frameworks able to connect multi-modal data views. In this context, here, by leveraging single cell data from a set of *IDH1/2* mutant AML patients, we develop deep learning approaches to explore how genotypic changes are reflected in cell specific gene expression signals. Specifically, we set out to answer if and how single cell gene expression patterns together with the

deployment of neural networks have the capacity to predict a cell's malignant status and genotype for specific genomic abnormalities.

Data & Methods

The study cohort consists of 4 healthy individuals and 6 AML patients, 3 with clonal *IDH1* and 3 with clonal *IDH2* mutations. These patients also harbored co-mutations in *NPM1*, *NRAS*, *KRAS*, *SRSF2*, *DNMT3A*, as well as a set of chromosomal abnormalities (gains in chromosomes 1q [+1q/dupli_chr1], 6 [+6/dupli_chr6], 8 [+8/dupli_chr8], 10 [+10/dupli_chr10] and 14 [+14/dupli_chr14]). For this cohort, scRNA-seq data were generated from BM and peripheral blood (PB) samples. Next to the single cell gene expression profiles, single cell genotypic information was also available for the 6 AML patients, as derived from the method of genotyping of transcriptomes (Nam et al. Nature 2019). (We note that the generation of the data is not part of the current thesis). After alignment and quality control, single cell gene expression counts for both AML and healthy individuals were generated, normalized and then integrated into a unified dataframe. The cells, from the AML patients, with at least one detected mutation or chromosomal abnormality were labeled as malignant while every cell from the WT individuals was labeled as WT. To assess if single cell gene expression values can predict the malignant or WT status of a cell we train a feedforward neural network that outputs the probability of a cell being malignant (binary classification model). Additionally, to be able to predict the acquired genomic abnormalities harbored by the malignant cells (for instance, if NRAS mutation is present or not), we train a multi-label classification model of a similar architecture. Application of holdout randomization tests (HRT, Tansey et al. JCGS 2021) in both trained models selects features predictive for the respective labels (malignant for the binary model, genetic abnormality present for the multi-label model).

Results

A total of 50,026 cells (35,314 were malignant and 14,712 WT) with high-quality data were selected for the training, validation and testing of the binary classification model. This binary model separated malignant and WT cells with an accuracy of 98%, precision of 98% and recall of 99%. Additionally, the HRT method led to the selection of 58 genes as important for this classification task (malignant vs WT). Gene ontology analysis on this set of 58 genes showed enrichment of processes related to apoptosis (Benjamini-Hochberg [BH] adjusted p-value = 0.009, e.g. *MCL1, HMGB2*) and the TGF-beta signaling pathway (BH adjusted p-value = 0.005, e.g. *ID1, JUNB*). Applying the model on the cells of the AML patients that were not part of the training, validation and test sets, revealed a small portion of cells within each patient that present a phenotype similar to that of the WT cells (WT-like). These WT-like predictions are the 4.1% of this cell-set and 56% of them correspond to myeloid differentiated cells. Similarly to the binary classification case, the multi-label model, trained, validated and tested on 16,614 cells of a single AML patient, presents 98% correct predictions in separating the malignant from the WT cells on the holdout test set. Additionally, this multi-label model achieved near optimal results on the prediction of the chromosomal abnormalities (AUC ROC higher >= 0.96) and had a considerable performance for the subclonal *NRAS* (AUC ROC = 0.83) mutation, in contrast to its limited capacity to correctly predict the mutational status of clonal *IDH1* mutation.

Discussion

This study develops deep learning approaches to explore how genotypic changes are reflected in cell specific gene expression signals in *IDH1/2* AML. The designed networks predict malignant vs WT cell status and identify the mutational status of specific genomic abnormalities for a single patient, while dealing concurrently with the excessive absence of mutation labels for some of these abnormalities during the training process. Both models showed similarly high performance in classifying malignant cells from WT ones. The low performance on predicting the *IDH1* status can be attributed to its low genotyping efficiency, while the notable performance on predicting the subclonal *NRAS* status implies the acquisition of specific gene expression profiles from the cells that acquire the *NRAS* mutation as a later event. This outcome demonstrates that the multi-label classification task may perform optimal when addressing cells with a representative spread of mutant and WT profiles such as subclones. This is of significant translational and clinical relevance as it is often such emerging subclones that carry mutations that confer resistance to treatment, and that seed disease relapse and progression. Lastly, the treating of both trained models as black boxes and the application of the HRT feature selection method, select as important input genes related to processes previously reported in the context of AML (e.g. apoptosis).

Conclusion

The analytical frameworks presented hereby demonstrate that the deployment of multi-view data integration concepts for the mining of bulk and single cell sequencing data in myeloid neoplasms, leads to a comprehensive and detailed profiling of molecular landscapes and enhances the capturing of genotype-phenotype associations. The derived outcomes show that these approaches offer the opportunity to establish connections between diverse data views and extract key signals related to disease biology. In a broader perspective, the rationale used for the integration of different data views from bulk RNA-seq data, can be applied to other studies investigating the role of splicing factor mutations across relevant signals that can be quantified by transcriptome sequencing (expression, splicing, transcript usage). Additionally, in other studies in oncology, especially in cancer indications with the presence of different genetic clones, the deployment of supervised deep learning architectures can link single cell transcriptomic and genomic data and show if and how the mutations and their clonality are reflected in the single cell gene expression profiles. Extending data views to include other types of diagnostic modalities such as morphological, immunophenotypic and clinical, as well as building integration approaches to analyze these views in a collective manner in supervised and unsupervised contexts, will pave the way for the adoption of the resulting insights in clinical practice.

Περίληψη

Οι μυελογενείς κακοήθειες αποτελούν ετερογενείς διαταραχές κλωνικών βλαστοκυττάρων, που οφείλονται σε γενετικές αλλοιώσεις και οδηγούν σε ελαττωματική αιμοποίηση. Η έρευνα των μηχανισμών των μυελογενών νεοπλασμάτων βασίζεται σε πειραματικά μοντέλα βιολογίας και στο φαινοτυπικό προσδιορισμό πρωτογενών δειγμάτων ασθενών μέσω ανερχόμενων γονιδιωματικών τεχνολογιών. Στο περιθώριο της αυξανόμενης πολυπλοκότητας, του όγκου και της διαστασιμότητας των δεδομένων που δημιουργούνται από αυτές τις πρακτικές, η εν λόγω διατριβή αναπτύσσει υπολογιστικά πλαίσια χρησιμοποιώντας διαφορετικά ομικά προφίλ (γονιδιωματικά, μεταγραφωματικά, επιγονιδιωματικά) και είδη τεχνικών αλληλουχίας (μαζικής και μεμονωμένων κυττάρων), με στόχο να ενισχύσει την κατανόηση της υποκείμενης βιολογίας των μυελογενών νεοπλασμάτων. Συγκεκριμένα, η εργασία αυτή στηρίζεται στη σύμπραξη και τη διασύνδεση πολλαπλών όψεων δεδομένων για την ολιστική ανάλυση σημάτων, αποσκοπώντας να αποσαφηνίσει μοριακά τοπία και να επικουρήσει τη μελέτη φαινότυπων σε γενετικό επίπεδο.

Το Κεφάλαιο 2 ερευνά τα τοπία μεταγραφώματος και χρωματίνης των *SF3B1* μεταλλαγμένων Μυελοδυσπλαστικών Συνδρόμων (MDS), αξιοποιώντας δεδομένα μαζικής αλληλουχίας RNA και ATAC από ισογονικές υγιείς και *SF3B1* μεταλλαγμένες σειρές επαγόμενων πολυδύναμων βλαστικών κυττάρων (iPSCs). Ειδικά, υλοποιούμε ένα αναλυτικό πλαίσιο συγχώνευσης πληροφοριών από διαφορετικά επίπεδα δεδομένων αλληλουχίας RNA (μάτισμα, χρήση μεταγραφημάτων, έκφραση γονιδίων). Το πλαίσιο αυτό ξεχωρίζει ένα σύνολο συμβάντων ματίσματος από 34 γονίδια, το οποίο σχετίζεται με την κατάσταση της μετάλλαξης *SF3B1* σε πρωτογενή δείγματα ασθενών με MDS. Παράλληλα, η ανάλυση της προσβασιμότητας της χρωματίνης δείχνει αυξημένη παρουσία μοτίβων TEAD σε ένα σύνολο ανοιχτών περιοχών της καθώς και αυξημένη κλίση των αιμοποιητικών προγονικών κυττάρων *SF3B1* προς την κατεύθυνση των μεγακαρυοκυττάρων - ερυθροειδών. Συνολικά, το κεφάλαιο αυτό, συνδυάζει προβολές από δεδομένα RNA για να κατηγοριοποιήσει γονιδιακούς στόχους με ελαττωματικό μάτισμα, ενώ επίσης παρέχει μια επισκόπηση του τοπίου της χρωματίνης *SF3B1* MDS και προτείνει προγράμματα μεταγραφής με πιθανούς ρόλους στη βιολογία της νόσου MDS.

Το Κεφάλαιο 3 εξετάζει εάν η έκφραση γονιδίων μεμονωμένων κυττάρων μαζί με την υπολογιστική ικανότητα των νευρωνικών δικτύων μπορούν να προβλέψουν χαρακτηριστικά του κυτταρικού γονότυπου στην *IDH1/2* μεταλλαγμένη Οξεία Μυελογενή Λευχαιμία (AML). Συγκεκριμένα, χρησιμοποιώντας δεδομένα αλληλουχίας RNA μεμονωμένων κυττάρων από 50.026 κύτταρα, εκπαιδεύτηκε ένα νευρωνικό δίκτυο που προβλέπει την κακοήθη ή υγιή κατάσταση του κυττάρου με ακρίβεια 98%. Στη συνέχεια, δοκιμάστηκε μια παρόμοια αρχιτεκτονική για την ταυτόχρονη πρόβλεψη συγκεκριμένων γονιδιωματικών ανωμαλιών σε μεμονωμένα κύτταρα, η οποία παρουσίασε macro-average AUC ROC=0.84 και AUC ROC=0.83 για την πρόβλεψη της μετάλλαξης *NRAS*. Εν κατακλείδι, το κεφάλαιο 3, μέσω επιβλεπόμενου deep learning, συνδέει τα προφίλ γονιδιακής έκφρασης και γονότυπου μεμονωμένων κυττάρων στην *IDH1/2* μεταλλαγμένη AML και δείχνει την προοπτική τέτοιων προσεγγίσεων μοντελοποίησης στην αποτύπωση σχέσεων γονότυπου-φαινότυπου.

Εκτενής περίληψη

Εισαγωγή

Τα μυελογενή νεοπλάσματα (MNs) αποτελούν ένα συνεχές φάσμα κλωνικών πολλαπλασιαστικών διαταραχών, που περιλαμβάνουν χρόνιες φάσεις, όπως τα Μυελοπολλαπλασιαστικά Νεοπλασματα (MPN) και τα Μυελοδυσπλαστικά Σύνδρομα (MDS) καθώς και οξείες φάσεις, όπως η Οξεία Μυελογενής Λευχαιμία (AML). Τα MNs συνήθως προκαλούνται από την απόκτηση γενετικών ανωμαλιών που διαταράσσουν τη φυσιολογική αιμοποίηση. Τα τελευταία χρόνια, οι μελέτες προφίλ γονιδιώματος παρέχουν ένα λεπτομερές κατάλογο των σωματικών μεταλλάξεων στα MNs. Ωστόσο, η κατανόηση των τελεστών της βιολογίας της νόσου και των μηχανισμών που οδηγούν στο μυελογενή μετασχηματισμό, βασίζεται στην ανάπτυξη πειραματικών μοντέλων (ποντικών, κυτταρικών, οργανοειδών) και στο φαινοτυπικό προσδιορισμό πρωτογενών δειγμάτων ασθενών. Τέτοιες προσεγγίσεις συνήθως συνδυάζονται με τεχνικές ανάλυσης του γονιδιώματος, του μεταγραφώματος και του επιγονιδιώματος του δείγματος. Παράλληλα με τις καθιερωμένες τεχνικές μαζικής αλληλουχίας, οι πιο πρόσφατες εξελίξεις στις τεχνολογίες μεμονωμένων κυττάρων συνεισφέρουν επίσης στην πλέον σύνηθη αποδοτική και εκτεταμένη παραγωγή ομικών συνόλων δεδομένων. Αυτά τα σύνολα δεδομένων περιλαμβάνουν διακριτές προβολές (αναπαραστάσεις ή σύνολα χαρακτηριστικών που προέρχονται από τα μετρούμενα βιομόρια) που επιτρέπουν τη μελέτη των μοριακών ιδιοτήτων σε πολλαπλά ομικά επίπεδα (γονιδιωματικό, μεταγραφικό ή επιγονιδιωματικό). Ωστόσο, οι αναλύσεις που επικεντρώνονται σε μια μόνο προβολή δεδομένων δεν οδηγούν στον πλήρη χαρακτηρισμό των μοριακών τοπίων και στην καθιέρωση των συσχετίσεων γονότυπου-φαινότυπου. Έτσι, η ανάπτυξη αναλυτικών πλαισίων που επιτρέπουν την ενοποίηση και την ερμηνεία πολλαπλών προβολών δεδομένων προσφέρει την ευκαιρία να μελετηθεί η αναπαράσταση διαφορετικών μοριακών επιπέδων καθώς και των σχέσεων μεταξύ τους. Λαμβάνοντας υπόψη την αυξανόμενη πολυπλοκότητα, κλίμακα και διαστασιμότητα των παραγόμενων δεδομένων, καθώς και την ανάγκη απόκτησης πολυδιάστατων ενδοσκοπήσεων σχετικά με τη συμπεριφορά της νόσου, αυτή η διατριβή αναπτύσσει αναλυτικά πλαισία με στόχο την ενίσχυση της κατανόησης της υποκείμενης βιολογίας των μυελογενών νεοπλασμάτων. Ειδικότερα, αυτή η εργασία χρησιμοποιεί αρχές από τη σύμπραξη και τη διασύνδεση πολλαπλών όψεων(προβολών) δεδομένων ως μέσο ενοποίησης σημάτων εντός και μεταξύ ομικών κατηγοριών (γονιδιωμα, μεταγράφωμα, επιγονιδίωμα) που προέρχονται από τεχνικές είτε μαζικής αλληλουχίας ή μεμονωμένων κυττάρων. Οι παρουσιαζόμενες αναλύσεις και τα πλαίσια που αναπτύχθηκαν στο πλαίσιο αυτής της διατριβής, αποσκοπούν στην αποσαφήνιση των μοριακών τοπίων και στην επικούριση της μελέτης φαινότυπων σε γενετικό επίπεδο, εστιάζοντας την προσοχή στα MDS με μετάλλαξη *SF3B1* και στην AML με μετάλλαξη *IDH1/2* αντίστοιχα.

Patient-specific MDS-RS iPSCs define the mis-spliced transcript repertoire and chromatin landscape of *SF3B1*-mutant HSPCs

Background

Τα μυελοδυσπλαστικά σύνδρομα (MDS) είναι μυελογενείς κακοήθειες που χαρακτηρίζονται από αναποτελεσματική αιμοποίηση, κυτταροπενίες αίματος και αυξημένο κίνδυνο εξέλιξης σε AML. Πρόσφατες μελέτες αλληλουχίας έχουν τονίσει τον ρόλο των μεταλλάξεων στα γονίδια του ματίσματος (*SF3B1*, *SRSF2*, *ZRSR2*, *U2AF1*) ως εναρκτήριο και καθοριστικό για τα MDS. Μεταξύ αυτών των γονιδίων, το *SF3B1* είναι το πιο συχνά μεταλλαγμένο στους ασθενείς με MDS (~ 24 %) και ορίζει μια ξεχωριστή νοσολογική οντότητα, που ονομάζεται MDS με δακτυλιοειδείς σιδεροβλάστες (MDS-RS). Οι μεταλλάξεις στο *SF3B1* εντοπίζονται συνήθως ως μεμονωμένα συμβάντα, στοχεύουν κυρίως στο hotspot K700 και σχετίζονται με ευνοϊκά προγνωστικά. Παρά τον χαρακτηρισμό του μοριακού τοπίου στο MDS, παραμένει ασαφές το πώς αυτές οι μεταλλάξεις οδηγούν την παθογένεση της νόσου και πώς μπορούν να ενημερώσουν την κλινική διαχείριση. Σε αυτή τη μελέτη, αξιοποιήσαμε δεδομένα από ένα πειραματικό μοντέλο επαγόμενων πολυδύναμων βλαστικών κυττάρων (iPSC) για να διερευνήσουμε τις συνέπειες της μετάλλαξης $SF3B1^{K700E}$ και τον ρόλο της στην παθογένεση της νόσου, ενσωματώνοντας πολλαπλές προβολές (όψεις) από το μεταγράφωμα και εξετάζοντας το τοπίο προσβασιμότητας της χρωματίνης.

Data & Methods

Για τη μελέτη αυτή, χρησιμοποιήσαμε ένα πάνελ 18 συνολικά $SF3B1^{K700E}$ μεταλλαγμένων και υγιών ($SF3B1^{WT}$) ισογονικών σειρών iPSC από 3 ασθενείς με MDS-RS που έφεραν μεμονωμένες μεταλλάξεις $SF3B1^{K700E}$. Για αυτές τις σειρές iPSC, πραγματοποιήθηκε κατευθυνόμενη αιμοποιητική διαφοροποίηση χρησιμοποιώντας πρωτόκολλα από το εργαστήριο της κας Παπαπέτρου και συλλέχθηκαν CD34+/CD45+ ανθρώπινα βλαστικά και προγονικά κύτταρα (HSPCs) για RNA και ATAC-sequencing. (Σημειώνουμε ότι η δημιουργία των δεδομένων δεν αποτελεί μέρος της τρέχουσας διατριβής). HSPCs από 16 σειρές iPSC συμπεριλήφθηκαν στις αναλύσεις RNA-seq μετά από ποιοτικό έλεγχο των αρχικών δεδομένων. Μετά την ευθυγράμμιση των αναγνώσεων RNA-seq ποσοτικόποιήθηκε η αφθονία των μεταγραφημάτων και η γονιδιακή έκφραση. Πραγματοποιήθηκαν αναλύσεις διαφορικής έκφρασης γονιδίου, διαφορικής χρήσης μεταγραφημάτων και διαφορικού ματίσματος (συναρμογής) ανάμεσα στα κύτταρα $SF3B1^{K700E}$ και στα κύτταρα $SF3B1^{WT}$. Για να αξιολογήσουμε τον αντίκτυπο της μετάλλαξης $SF3B1^{K700E}$ σε επίπεδο εξονίου, μεταγραφήματος και γονιδίου, συνδυάσαμε σήματα από αυτές τις 3 αναλύσεις μέσω ενός πολυσταδιακού πλαισίου συγχώνευσης. Αρχικά, εντοπίσαμε το σύνολο των μεταγραφημάτων που περιέχουν τα εξόνια που συμμετέχουν σε συμβάντα διαφορικού ματίσματος. Στη συνέχεια, συσχετίσαμε κάθε συμβάν διαφορικού ματίσματος με το σύνολο των διαφορικά χρησιμοποιούμενων μεταγραφημάτων. Τα ζεύγη που ανήκαν σε γονίδια με στατιστικά σημαντική διαφορική έκφραση και περιείχαν ένα διαφορικό συμβάν ματίσματος με τιμή FDR (False Discovery Rate) εντός των 20 χαμηλότερων, θεωρήθηκαν ως το σύνολο «βαθμίδας 1». Αυτό το σύνολο, περιέχει τα μεταλλαγμένα συμβάντα και γονίδια της υπογραφής *SF3B1*. Επιπλέον, κύτταρα HSPC από 15 σειρές iPSC συμπεριλήφθηκαν στην ανάλυση ATAC-seq. Μετά την ευθυγράμμιση των αναγνώσεων και τον ποιοτικό έλεγχο, εντοπίσαμε κορυφές προσβασιμότητας χρωματίνης και δημιουργήσαμε έναν άτλαντα ATAC-seq. Αυτός ο άτλας χρησιμοποιήθηκε για ανάλυση διαφορικής προσβασιμότητας χρωματίνης, συσχέτιση με το τοπίο προσβασιμότητας της φυσιολογικής αιμοποιητικής ιεραρχίας καθώς και για ανάλυση εμπλουτισμού μοτίβων.

Results

Η ανάλυση κύριων συνιστωσών (PCA) και η ιεραρχική ομαδοποίηση (hierarchical clustering) με βάση τη γονιδιακή έκφραση ομαδοποίησαν τις σειρές iPSC ανά γονότυπο (*SF3B1*^{K700E} vs *SF3B1*^{WT}). Επιπλέον, διαφορικές αναλύσεις αποκάλυψαν 2737 διαφορικά εκφραζόμενα γονίδια, 1086 διαφορικά χρησιμοποιούμενα μεταγραφήματα και 1829 διαφορικά συμβάντα ματίσματος (συναρμογής) μεταξύ των κυττάρων *SF3B1*^{K700E} και *SF3B1*^{WT}. Η ενοποίηση των σημάτων από αυτές τις αναλύσεις χρησιμοποιώντας πολυσταδιακή προσέγγιση συγχώνευσης είχε ως αποτέλεσμα την παραγωγή μιας υπογραφής ματίσματος που αποτελείται από 59 συμβάντα (ματίσματος) προερχόμενα από 34 γονίδια. Δοκιμάσαμε αυτήν την υπογραφή σε ένα δημοσιευμένο σύνολο δεδομένων από πρωτογενή δείγματα ασθενών με MDS (Pellagatti et al. Blood 2018). Συγκεκριμένα, PCA με βάση το επίπεδο συμπερίληψης των συμβάντων συναρμογής της υπογραφής μας, διαχώρισε τους ασθενείς με μετάλλαξη *SF3B1* από τους ασθενείς χωρίς μεταλλάξεις σε παράγοντες ματίσματος (SF-WT) ή τα υγιή άτομα. Σημαντικότερα, εντόπισε έναν ασθενή που ομαδοποιήθηκε μαζί με τους ασθενείς με μετάλλαξη *SF3B1* ενώ εσφαλμένα είχε σημειωθεί ως SF-WT. Αυτός ο ασθενής είχε μια προηγουμένως παραβλεφθείσα διαγραφή 6 bp στο σημείο K700E. Συγκρίνοντας τον άτλαντα κορυφών ATAC-seq με τα προφίλ προσβασιμότητας χρωματίνης των πρωτογενών τύπων των ανθρώπινων κυττάρων κατά μήκος της αιμοποιητικής ιεραρχίας (Corces et al. Nat Genetics 2016), διαπιστώσαμε ότι το τοπίο χρωματίνης των *SF3B1*^{K700E} HSPCs έμοιαζε περισσότερο με αυτό των μεγακαρυωτικών-ερυθροειδών προγονικών κυττάρων (MEPs) και των ερυθροειδών (Ery) κυττάρων. Επιπλέον, το σύνολο εμπλουτισμένων μοτίβων στις κορυφές ATAC-Seq που ήταν πιο προσβάσιμες σε κύτταρα *SF3B1*^{K700E} και επιπλέον συνδέονταν με γονίδια αυξημένης έκφρασης στα κύτταρα αυτά (*SF3B1*^{K700E}), περιελάμβαναν μοτίβα της οικογένειας μεταγραφικών παραγόντων TEAD. Τα *TEAD2* και *TEAD4* είχαν αυξημένη έκφραση στα κύτταρα *SF3B1*^{K700E} σε σύγκριση με τα *SF3B1*^{WT}. Ακόμα, η μεταγραφική δραστηριότητα TEAD, μετρήμένη με μία κατασκευή αναφοράς λουσιφεράσης, ήταν υψηλότερη στα κύτταρα *SF3B1*^{K700E} σε σύγκριση με τα *SF3B1*^{WT}. Επίσης, δε βρήκαμε έκφραση ή ενεργοποίηση του YAP ή TAZ, τα οποία συνδέονται με το DNA ως σύμπλεγμα με το TEAD κατά την ενεργοποίηση της οδού σηματοδότησης Hippo. Τα στοιχεία αυτά υποδηλώνουν αυξημένη έκφραση και δραστηριότητα του TEAD ανεξάρτητα από την οδό σηματοδότησης Hippo στα *SF3B1*^{K700E} κύτταρα.

Discussion

Με την υποστήριξη ενός πλαισίου ενοποίησης (integration) δεδομένων, αυτή η μελέτη αποτιμά συνδυαστικά τις επιδράσεις της μετάλλαξης *SF3B1*^{K700E} σε παράλληλα επίπεδα απορρύθμισης του μεταγραφώματος με στόχο την ταξινόμηση των συμβάντων ματίσματος σε βαθμίδες. Συγκεκριμένα, αυτό το πλαίσιο αξιολογεί συστηματικά τις σχέσεις μεταξύ της μετάλλαξης *SF3B1*, του διαφορικού ματίσματος, της χρήσης μεταγραφημάτων και της έκφρασης γονιδίων και καταλήγει σε μια χαρακτηριστική υπογραφή ματίσματος *SF3B1*^{K700E}. Αυτή η υπογραφή περιλαμβάνει πολλά γνωστά γονίδια και είναι σε θέση να αναγνωρίσει ατυπικές μεταλλάξεις που αφορούν το hotspot K700. Επιπλέον, αυτή η μελέτη δείχνει, σε επίπεδο χρωματίνης, ένα δυνητικό προσανατολισμό των HSPCs *SF3B1*^{K700E} κυττάρων προς την ερυθροειδή κατεύθυνση σε σχέση με τη μυελοειδή - ένα στοιχείο που μπορεί να σχετίζεται με την προτιμητέα συμμετοχή της ερυθροειδούς κατεύθυνσης (lineage) στα MDS και ειδικότερα στα MDS-RS. Τέλος, οι αναλύσεις προσβασιμότητας της χρωματίνης υποστηρίζουν έναν θεωρούμενο ρόλο των παραγόντων

μεταγραφής TEAD στο πλαίσιο της μετάλλαξης *SF3B1*^{K700E}, ένα σήμα που χρήζει επικύρωσης σε μελλοντικές μελέτες.

Predicting single cell genotypes from single cell expression profiles in AML using deep learning

Background

Περίπου το 30% των περιπτώσεων με MDS τελικά εξελίσσονται σε AML, έναν επιθετικό καρκίνο του αίματος που εκδηλώνεται ταχέως, έχει αδύναμη ανταπόκριση στη θεραπεία και δυσοίωνα προγνωστικά επιβίωσης. Η AML είναι μια γενετικά ετερογενής νόσος που ορίζεται από τη σταδιακή συσσώρευση μεταλλάξεων. Αυτές συχνά χαρακτηρίζονται από συγκεκριμένες αλληλεπιδράσεις γονιδίων, που υποδηλώνουν λειτουργική συνεργασία, και οδηγούν σε γενετικά και κλωνικά ετερογενείς πληθυσμούς. Αυτό θέτει μια σημαντική πρόκληση στην αντιμετώπιση και τελικά τη θεραπεία της νόσου. Η αποσαφήνιση του ρόλου και της επίδρασης αυτής της ποικιλομορφίας στους κυτταρικούς φαινότυπους και τη βιολογία της νόσου απαιτεί: 1. μοριακές αναπαραστάσεις σε κυτταρικό επίπεδο, οι οποίες δεν μπορούν να επιτευχθούν με προσεγγίσεις μαζικής αλληλουχίας σε πρωτογενή δείγματα, και 2. αναλυτικά πλαίσια ικανά να συνδέσουν πολυτροπικές προβολές δεδομένων. Στο περιθώριο αυτό, αξιοποιώντας δεδομένα μεμονωμένων κυττάρων από ένα σύνολο ασθενών AML με μετάλλαξη *IDH1/2,* αναπτύσσουμε προσεγγίσεις deep learning για να διερευνήσουμε πώς οι γονοτυπικές αλλαγές μεμονωμένων κυττάρων αντανακλώνται στα σήματα γονιδιακής έκφρασης. Συγκεκριμένα, αποσκοπούμε να απαντήσουμε εάν και πώς τα μοτίβα γονιδιακής έκφρασης ενός κυττάρου σε συνδυασμό με την χρήση νευρωνικών δικτύων, έχουν τη δυνατότητα να προβλέψουν την κακοήθη κατάσταση και τον γονότυπο ενός κυττάρου για συγκεκριμένες γονιδιωματικές ανωμαλίες.

Data & Methods

Το σύνολο των δεδομένων της μελέτης αυτής προέρχεται από 4 υγιή άτομα και 6 ασθενείς με AML, 3 με κλωνικές μεταλλάξεις *IDH1* και 3 με κλωνικές μεταλλάξεις *IDH2*. Αυτοί οι ασθενείς είχαν επίσης μεταλλάξεις στα γονίδια *NPM1*, *NRAS*, *KRAS*, *SRSF2*, *DNMT3A*, καθώς και ένα σύνολο χρωμοσωμικών ανωμαλιών (αυξήσεις [gains] στα χρωμοσώματα 1q [+1q/dupli_chr1], 6 [+6/dupli_chr6], 8 [+8/dupli_chr8], 10 [+10/dupli_chr10] και 14 [+14/dupli_chr14]). Για αυτό το σύνολο ασθενών και υγιών ατόμων, δεδομένα scRNA-seq δημιουργήθηκαν από δείγματα μυελού των οστών (Bone Marrow) και περιφερικού αίματος (Peripheral Blood). Παράλληλα με τα προφίλ έκφρασης γονιδίων από μεμονωμένα κύτταρα, ήταν επίσης διαθέσιμες γονοτυπικές πληροφορίες μεμονωμένων κυττάρων για τους 6 ασθενείς με AML, όπως προέκυψαν από τη μέθοδο GoT (Nam et al. Nature 2019). (Σημειώνουμε ότι η δημιουργία των δεδομένων δεν αποτελεί μέρος της τρέχουσας διατριβής). Μετά την ευθυγράμμιση των αναγνώσεων και τον ποιοτικό έλεγχο των δεδομένων, υπολογίστηκαν, κανονικοποιήθηκαν και στη συνέχεια ενσωματώθηκαν σε ένα ενιαίο πίνακα οι γονιδιακές εκφράσεις μεμονωμένων κυττάρων τόσο για τους ασθενείς όσο και για υγιή άτομα. Τα κύτταρα, από τους ασθενείς με AML, τα οποία είχαν τουλάχιστον μία ανιχνευμένη μετάλλαξη ή

χρωμοσωμική ανωμαλία επισημάνθηκαν ως κακοήθη ενώ κάθε κύτταρο από τα υγιή άτομα επισημάνθηκε ως υγιές. Για να εκτιμήσουμε εάν οι τιμές έκφρασης ενός γονιδίου ενός κυττάρου μπορούν να προβλέψουν την κακοήθη ή υγιή κατάσταση ενός κυττάρου, εκπαιδεύουμε ένα feedforward νευρωνικό δίκτυο που εξάγει την πιθανότητα ένα κύτταρο να είναι κακοήθες (μοντέλο δυαδικής ταξινόμησης). Επιπλέον, για να είμαστε σε θέση να προβλέψουμε τις γονιδιωματικές ανωμαλίες που φέρουν τα κακοήθη κύτταρα (για παράδειγμα, εάν υπάρχει μετάλλαξη *NRAS* ή όχι), εκπαιδεύουμε ένα μοντέλο πολλαπλών ετικετών (multi-label) παρόμοιας αρχιτεκτονικής. Η εφαρμογή της μεθόδου HRT (Holdout Randdomization Test, Tansey et al. JCGS 2021) και στα δύο εκπαιδευμένα μοντέλα επιλέγει τα χαρακτηριστικά με την κυριότερη δυνατότητα πρόβλεψης των αντίστοιχες ετικετών (κακοήθη κύτταρα για το δυαδικό μοντέλο, γενετικές ανωμαλίες για το μοντέλο πολλαπλών ετικετών).

Results

Συνολικά 50.026 κύτταρα (35.314 κακοήθη και 14.712 υγιή) με δεδομένα υψηλής ποιότητας επιλέχθηκαν για την εκπαίδευση, την επικύρωση και τη δοκιμή του μοντέλου δυαδικής ταξινόμησης. Αυτό το δυαδικό μοντέλο διαχώρισε τα κακοήθη από τα υγιή κύτταρα με accuracy 98%, precision 98% και recall 99%, ενώ η μέθοδος HRT οδήγησε στην επιλογή 58 γονιδίων ως σημαντικών για αυτήν την ταξινόμηση (κακοήθη κύτταρα έναντι υγιών). Η ανάλυση γονιδιακής οντολογίας σε αυτό το σύνολο των 58 γονιδίων έδειξε τη συμμετοχή διεργασιών που σχετίζονται με την απόπτωση (BH adjusted p-value = 0,009, πχ *MCL1, HMGB2*) και την οδό σηματοδότησης TGF-beta (BH adjusted p-value = 0,005, πχ *ID1, JUNB*). Εφαρμόζοντας το μοντέλο αυτό στα κύτταρα των ασθενών με AML που δεν αποτελούσαν μέρος των σετ εκπαίδευσης, επικύρωσης και δοκιμών, αποκάλυψε ένα μικρό ποσοστό (κυττάρων) σε κάθε ασθενή που παρουσίαζε ένα φαινότυπο παρόμοιο με αυτόν των υγιών κυττάρων (WT-like). Οι εν-λόγω WT-like προβλέψεις είναι το 4,1% του συνόλου και το 56% αυτών αντιστοιχούν σε μυελογενή διαφοροποιημένα κύτταρα. Παρομοίως με την περίπτωση δυαδικής ταξινόμησης, το μοντέλο πολλαπλών ετικετών, εκπαιδευμένο, επικυρωμένο και δοκιμασμένο σε 16.614 κύτταρα ενός ασθενούς με AML, παρουσιάζει 98% σωστές προβλέψεις κατά τον διαχωρισμό των κακοήθων κυττάρων από τα υγιή στο σετ δοκιμής. Επιπλέον, αυτό το μοντέλο πολλαπλών ετικετών πέτυχε σχεδόν βέλτιστα αποτελέσματα στην πρόβλεψη των χρωμοσωμικών ανωμαλιών (AUC ROC >= 0,96) και είχε σημαντική απόδοση για την πρόβλεψη της υποκλωνικής μετάλλαξης *NRAS* (AUC ROC = 0,83), σε αντίθεση με την περιορισμένη ικανότητά του να προβλέπει σωστά την κατάσταση της κλωνικής μετάλλαξης *IDH1*.

Discussion

Αυτή η μελέτη αναπτύσσει προσεγγίσεις deep learning για να διερευνήσει πώς οι γονοτυπικές αλλαγές αντανακλώνται στα σήματα γονιδιακής έκφρασης μεμονωμένων κυττάρων στην AML με μεταλλάξεις *IDH1/2*. Τα σχεδιασμένα δίκτυα προβλέπουν την κατάσταση κακοήθων έναντι υγιών κυττάρων και προσδιορίζουν την κατάσταση μετάλλαξης συγκεκριμένων γονιδιωματικών ανωμαλιών, αντιμετωπίζοντας ταυτόχρονα την απουσία ετικετών για ορισμένες από αυτές τις ανωμαλίες κατά τη διάρκεια της εκπαίδευσης. Και τα δύο μοντέλα έδειξαν εξίσου υψηλή απόδοση στην ταξινόμηση των κακοήθων κυττάρων έναντι των υγιών. Η χαμηλή απόδοση στην πρόβλεψη της κατάστασης *IDH1* μπορεί να αποδοθεί στη χαμηλή αποτελεσματικότητα της μεθόδου GoT κατά την παραγωγή των γονοτυπικών προφίλ, ενώ η

σημαντική απόδοση στην πρόβλεψη της κατάστασης *NRAS* υποδηλώνει την απόκτηση συγκεκριμένων προφίλ γονιδιακής έκφρασης από τα κύτταρα που την αποκτούν (μετάλλαξη *NRAS*) ως μεταγενέστερο γεγονός. Αυτό το αποτέλεσμα δείχνει ότι η ταξινόμηση πολλαπλών ετικετών για την πρόβλεψη μεταλλάξεων, μπορεί να έχει τη βέλτιστη απόδοση σε πληθυσμούς κυττάρων με ένα αντιπροσωπευτικό εύρος μεταλλαγμένων και υγιών προφίλ, όπως οι υποκλώνοι. Αυτό το στοιχείο εχει έννοια κλινικού χαρακτήρα καθώς και σημασία στο πλάισιο μεταφραστικής έρευνας, καθώς συχνά τέτοιοι αναδυόμενοι υποκλώνοι φέρουν μεταλλάξεις που προσδίδουν αντίσταση στη θεραπεία της νόσου αλλά και συνεισφέρουν στην υποτροπή της (νόσου). Τέλος, η διαχείριση και των δύο εκπαιδευμένων μοντέλων ως "μαύρων κουτιών" και η εφαρμογή της μεθόδου HRT επιλέγουν ως σημαντικά, γονίδια εισόδου που σχετίζονται με διεργασίες που έχουν επισημανθεί στη βιβλιογραφία της νόσου AML (πχ απόπτωση).

Συμπεράσματα

Τα αναλυτικά πλαίσια που παρουσιάζονται στην παρούσα διατριβή δείχνουν ότι η εφαρμογή εννοιών από το πεδίο ενοποίησης δεδομένων πολλαπλών προβολών κατά τη θεώρηση (mining) δεδομένων μαζικής αλληλουχίας αλλά και μεμονωμένων κυττάρων, οδηγεί σε μια περιεκτική και λεπτομερή αποτύπωση των μοριακών τοπίων και ενισχύει την καταγραφή των σχέσεων γονότυπου-φαινότυπου στα μυελογενή νεοπλάσματα. Τα αποτελέσματα που προκύπτουν υποδηλώνουν ότι αυτές οι προσεγγίσεις έχουν τη δυναμική να δημιουργήσουν συνδέσεις μεταξύ διαφορετικών όψεων δεδομένων και να εντοπίσουν σήματα που σχετίζονται με τη βιολογία της νόσου. Υπό ένα ευρύτερο πρίσμα, η λογική που χρησιμοποιήθηκε για την ενοποίηση προβολών από δεδομένα RNA-seq, μπορεί να εφαρμοστεί σε άλλες μελέτες που εξετάζουν το ρόλο των μεταλλάξεων παραγόντων ματίσματος σε σήματα που που μπορούν να ποσοτικοποιηθούν με αλληλουχία μεταγραφώματος (έκφραση γονιδίων, μάτισμα, χρήση μεταγραφημάτων). Επιπλέον, σε άλλες μελέτες στην ογκολογία και ειδικά σε περιπτώσεις καρκίνου με την παρουσία διαφορετικών γενετικών κλώνων, η ανάπτυξη εποπτευόμενων αρχιτεκτονικών deep learning μπορεί να συσχετίσει γονιδιωματικά και μεταγραφικά προφίλ μεμονωμένων κυττάρων και να δείξει εάν και πώς οι μεταλλάξεις και ο τύπος τους αντικατοπρίζονται στο προφίλ έκφρασης γονιδίων ενός κυττάρου. Η επέκταση των προβολών (όψεων) δεδομένων για τη συμπερίληψη και άλλων τύπων διαγνωστικών εξετάσεων όπως μορφολογικές, ανοσοφαινοτυπικές και κλινικές, καθώς και η ανάπτυξη ενοποιητικών (integrative) υπολογιστικών προσεγγίσεων για την ανάλυση αυτών των δεδομένων με συλλογικό τρόπο σε εποπτευόμενα και μη εποπτευόμενα περιβάλλοντα, θα ανοίξει το δρόμο για την υιοθέτηση των εξαξόμενων γνώσεων στην κλινική πράξη.

Table of contents

List of Figures

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 1

Introduction

1.1. Cancer: Pathogenesis and evolution

Cancer arises from the acquisition of mutations that result in the autonomous growth and expansion of malignant clones[1,2]. Somatic mutations occur spontaneously during lifetime and are largely inconsequential. However, a subset of these mutations may alter important biological processes that confer a fitness advantage to the carrier cells and contribute to malignant transformation[2,3]. The latter mutations, called drivers, enable the cell to escape the normal constraints of development and proliferation, contribute to the modification of key cellular functions and pathways, cause disorders and phenotypic alterations and lead to tumor formation[1,3].

Cancers are considered to share a common framework of pathogenesis and progression[3,4] whereby each tumor is the product of a Darwinian evolutionary process happening among the population of cells residing in the tissue microenvironments[4,5]. These microenvironments, shaped by the tissue space, resources, immune predation as well as a mixture of adverse conditions such as hypoxia and acidosis, pose constraints in tumor growth[5]. Similarly to the Darwinian principles of the evolution of species, cancer progression is driven by the stepwise accrual of mutations and the concomitant natural selection sweeps occurring on the resulting phenotypic diversity[3,4,6]. This process may confer a selective advantage to cells carrying genetic alterations that favor proliferation and survival and lead to the dominance of specific subpopulations. Genomic profiling of a tumor's DNA reveals the set of genetic variants accumulated during disease evolution. Quantitative estimates of the cellular representation for each mutation using variant allele fraction (VAF) metrics, provide further insights into the temporal order of mutation acquisition. This temporal order of mutations enables the reconstruction of the evolutionary tree of the tumor and sheds light on intra-tumor heterogeneity (ITH). Alterations present in all cells form the trunk of a cancer's somatic evolutionary trajectory while mutations identified in specific cell subsets define distinct subclones that arise and grow during later stages of disease evolution[5,6].

Cancer genome profiling studies[7,8] reveal a diverse spectrum of gene mutations, involving frequently mutated genes (>5% within tumor indications) as well as a long tail of infrequently mutated genes (<2% within tumor indications). On average each patient has 4 such driver mutations (ranging from 1-10 across tumor types)[2], resulting in a diverse range of patient-specific molecular profiles. This genetic heterogeneity underscores the complexity of modeling and treating cancer. Beyond gene mutations, the occurrence of

distinct epigenetic changes, including DNA methylation, chromatin remodeling and post-translational histone modifications also adds to the variety of resulting phenotypic patterns[5,6]. Therefore, ITH combined with the diversity between cancer patients outline the degree of the disease complexity and the challenges of patient care. With the advent of computational methods and technological pipelines such as next generation sequencing, it is now possible to map out changes in a tumor's genome, transcriptome and epigenome. Analysis of genotype-phenotype relationships also enables insights into mechanisms of disease biology. The generation of detailed and extensive datasets coupled with the design of quantitative approaches to study associations between genotypes, biological readouts and clinical phenotypes, significantly advance our potential to deliver evidence-based care and therapeutics in oncology[5,9].

The work in this thesis takes into consideration patient-relevant genetic and clonal representations in myeloid neoplasms, employs high-dimensional omics profiling applications (genetic, transcriptomic and epigenetic) and develops analytical frameworks to study how specific gene mutations contribute to disease pathogenesis in the context of myeloid neoplasms.

1.2. Hematopoietic System

Hematologic malignancies describe a heterogeneous set of myeloid and lymphoid neoplasms emerging from the disruption of normal hematopoiesis[10–12]. Hematopoiesis is the hierarchical developmental process of the lifelong and continuous formation of blood cells produced from a limited population of hematopoietic stem cells (HSCs)[11,13–16]. HSCs reside at the apex of the hematopoietic system and have the ability of self-renewal as well as of multipotent differentiation to all blood cell lineages[11,13–16]. HSCs are located in the bone marrow (BM) niche and in the absence of stimuli reside in a quiescent state of low mitochondrial activity and limited levels of protein synthesis[17]. In the context of maintaining homeostatic balance within the tissue, the regulation between the self-renewal and differentiation of HSCs is complex and dynamic and depends both on intracellular characteristics as well as extrinsic signals from the microenvironment[11,12]. Upon self-renewal HSCs produce new stem cells assisting in the supply of the HSC pool, while upon differentiation HSCs yield a variety of hematopoietic progenitor cells that gradually commit to specific lineages (myeloid and lymphoid) and progressively give rise to the mature blood cell types[11,13–16].

The hematopoietic process follows a hierarchical structure and unfolds as a continuum of multipotent hematopoietic stem and progenitor cells (HSPCs) that give rise to specialized blood cell lineages (Figure 1.1) in response to a regulated environment of growth factors and cellular interactions[11,12,15,16,18]. The classic roadmap of this tree-like structure is characterized by the commitment of progenitor intermediate cells towards one of two principal hematopoietic branches: the myeloid and the lymphoid[13,16,18] (Figure 1.1). The mature blood cells of the myeloid branch descend mostly from common myeloid progenitor cells (CMPs) and include monocytes, erythrocytes, granulocytes (neutrophils, eosinophils, mast cells, basophils) and megakaryocytes (Figure 1.1)[12,16,18]. On the other hand, the lymphoid branch generates B cells, natural killer (NK) cells as well as T cells[12,13,16]. This set of terminally differentiated blood cells called lymphocytes form the backbone of the immune system responding to a range of pathological challenges.

This thesis focuses specifically on blood cancers that arise following the deregulation of the myeloid lineage.

1.3. Myeloid Neoplasms

Myeloid neoplasms (MNs) are prevalent and clonal hematologic malignancies marked by the dysregulated proliferation and differentiation of HSCs and myeloid progenitor cells[19,20]. Genetic changes and epigenetic variations in these cells lead to abnormal growth and defective maturation of the myeloid cell types[21]. Diagnosis and classification of the MNs to separate clinicopathological disease entities relies heavily on the assessment of BM morphology, immunophenotyping, clinical features (such as the enumeration of peripheral blood counts [PBCs]), cytogenetic and gene mutation profiling[21]. MNs consist of chronic disorders like myelodysplastic syndromes (MDS) and myeloproliferative neoplasms (MPN), acute stages such as acute myeloid leukemia as well as related overlap syndromes, i.e. MDS/MPN. Apart from these disease groups, the umbrella of MNs, as per WHO 2022, also covers: myeloid precursor lesions, mastocytosis, secondary myeloid neoplasms, myeloid/lymphoid neoplasms with eosinophilia and defining gene rearrangement and acute leukemias of mixed or ambiguous lineage[22]. Even though each of these disorders presents with distinct clinical features, MNs are closely related, lie on a continuum of morphological parameters from dysplastic to more proliferative, share genetic features and as a result may have common therapeutic approaches. Treatment modalities include hypomethylating agents, hematopoietic stem cell transplantation, chemotherapy and more recently targeted agents.

The starting point of the pathophysiological process that ultimately induces MDS is the growth and expansion of a somatically mutated clone of hematopoietic cells[23,24]. In particular, the stage before disease presentation, called clonal hematopoiesis, begins with an initiating driver mutation in HSPCs. Clonal hematopoiesis manifests under selective forces induced by various exposures such as cytotoxic treatments and tobacco smoking, unrepaired DNA replication errors, aging or natural selection[24]. While in the setting of clonal hematopoiesis a single mutation can lead to clonal expansions, secondary co-operative mutations are required to confer a hematologic malignancy. Thus, the majority of people with clonal hematopoiesis may never transform into a blood cancerous stage and remain in a phase of ‘indeterminate potential’[24]. Therefore, this condition, named clonal hematopoiesis of indeterminate potential (CHIP), describes the clonal expansions of somatically mutated HSPCs in individuals with absence of dysplasias or cytopenias or any other diagnostic criteria for hematologic neoplasms[21,23–25]. However, upon the acquisition of secondary mutations, clonal hematopoiesis can dominate the BM and lead to malignant transformation and overt disease presentation. This malignant transformation, depending on the morphologic abnormalities as well as the cytogenetic and mutational landscape, can meet the criteria for an MN diagnosis[23,26]. MNs such as MDS or MPN can further progress to AML upon the increase of the abnormal immature blood cells called blasts, above the threshold of 20% [23].

Chapter 2 of the present thesis studies one of the most common mutations in MDS (*SF3B1*), whilst Chapter 3 is set in the context of *IDH* mutations in AML.

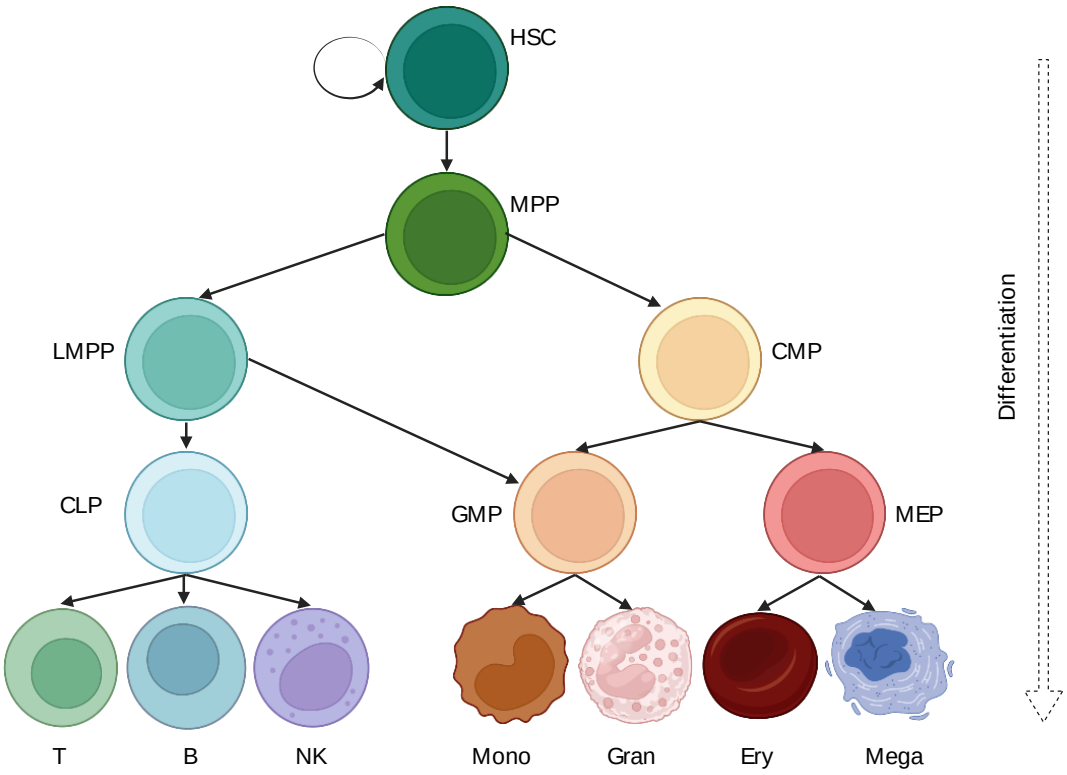


Figure 1.1. Hierarchical representation of the hematopoietic process adapted from Corces et al. 2016. Hematopoietic stem cells (HSCc) can either self-renew or produce multipotent progenitor cells (MPPs) that further differentiate to lymphoid-primed multipotent progenitor cells (LMPPs) and common myeloid progenitors (CMPs). The latter can further branch into megakaryocyte–erythroid progenitor cells (MEPs) or granulocyte-monocyte progenitor cells (GMPs) which can be also reached from LMPPs. The terminally differentiated set of myeloid cell types consists of monocytes (Mono), granulocytes (Gran), erythrocytes (Ery) and megakaryocytes (Mega). On the lymphoid branch, LMPPs give rise to common lymphoid progenitors (CLPs) that mature into B cells, natural killer (NK) cells as well as T cells. Drawn using BioRender.

1.3.1 Myelodysplastic Syndromes

Myelodysplastic syndromes (MDS) are a heterogeneous group of myeloid neoplasms characterized by dysplasia, ineffective hematopoiesis, varying cytopenias and a significant risk of transformation to AML[23,26–29]. MDS diagnosis and subtype classification are mostly based on the examination of morphological features (e.g. the degree of myelodysplasia, the presence of ring sideroblasts), percentage of bone marrow and peripheral blood blasts, specific cytogenetic abnormalities (5q deletion, monosomy 7) and gene mutations (*SF3B1*, *TP53*)[21,22,27,30]. The median age at diagnosis is approximately 70 years old[26,29].

Determining the genetic landscape of MDS is important for improving the diagnostic, therapeutic as well as prognostic practices in oncology[23,25,30–32]. The mutational burden as well as the detection of clonal and subclonal mutations have significant prognostic value[31,33]. Several of the genes that are mutated in MDS patients are commonly identified in the genetic profiling of other MNs (like MPN, MDS/MPN, AML) while

the complex patterns of co-mutations and subclonal evolution are associated with diverse disease trajectories and clinical traits. Upon the diagnosis of MDS, patients have a median number of 2-3 mutations while the landscape of the disease is characterized by mutations in more than 40 genes. However, only a small fraction of these are frequently mutated whereas the rest compose a long tail of more rare abnormalities[23,25,34]. Recurrently mutated genes in MDS include those involved in RNA splicing (*SF3B1*, *U2AF1*, *SRSF2* and *ZRSR2*), DNA methylation (*DNMT3A*, *TET2*, *IDH1*, *IDH2*), chromatin modification (*EZH2*, *ASXL1*, *KMT2*, *SUZ12*), transcription regulation (*TP53*, *EVI1*, *RUNX1*, *GATA2*), DNA repair control (*ATM*, *BRCC3*), and the cohesin complex (*STAG2*, *CTCF*, *SMC1A*, *RAD21*) [23,25,27,32,34]. Different combinations of mutations in these genes lead to the dysregulation of various biological pathways accounting for the disease heterogeneity among MDS patients. Events in RNA splicing, DNA methylation and histone modification genes are early driver and clonal events, while the rest contribute to disease evolution[25,27].

The most frequent mutations in MDS target components of the spliceosome machinery. Mutations in splicing factor genes are recurrent events and have been described as central to MDS disease biology[23,25,27,32,34]. Such mutations are heterozygous (e.g. affecting only one allele) and mutually exclusive (like *SF3B1* and *SRSF2*) and the different co-mutation patterns affect the downstream genomic evolution[27]. Given their importance in MDS pathogenesis, splicing factors have been subject to rapid therapeutic drug targeting, while inhibitors of the spliceosome complex are currently under clinical trial development[23,32]. The *SF3B1* gene is the most prevalently mutated one in MDS (~ 24% of the patients) and encodes the splicing factor 3b subunit 1[25,27,32]. Its somatic mutation is an early, disease-defining genetic lesion with an overall median variant allele frequency (VAF) around 40%[23,32]. Less frequently, *SF3B1* mutations can be identified as secondary events appearing in the background, most commonly, of *TET2*, *DNMT3A* or *ASXL1*-mutated cases[32]. From a clinicopathological perspective, *SF3B1* defines a distinct nosologic group in MDS, is associated with more favorable outcomes and has proven to have notable importance in prognostic systems of risk of transformation to AML [31,32,35]. *SF3B1*-mutant MDS is characterized by the presence of ring sideroblasts (RS), ineffective erythropoiesis, low blast counts and macrocytic anemia[21,23,32,35].

SF3B1 is a key component of the U2 small nuclear ribonucleoprotein complex (snRNP). Functionally, mutations in *SF3B1* affect splicing, namely the regulatory process that removes the intronic portions from the pre-mRNA and then ligates the protein-coding sequences (exons) of the genes in the context of transcription [36,37]. Under normal conditions, such exons are joined in different combinations (alternative splicing, Figure 1.2), resulting in a range of alternative transcripts. The acquisition of mutations in *SF3B1* (and other spliceosome genes) induces preferential splicing alterations, leading to differential splicing behavior compared to wild type [36,37]. *SF3B1* mutations cause alternative 3' splicing, mainly affecting mitochondrial gene pathways [35]. Specifically, such mutations alter the RNA branchpoint recognition leading to the preferential use of cryptic 3' splice sites (Figure 1.2). This results in decreased production of canonical transcripts and subsequent downregulation of protein expression as well as in increased formation of aberrant transcripts with a premature stop codon. The latter transcripts are degraded by nonsense-mediated decay mechanisms [23,32,38]. Apart from alternative 3' splice site events (A3SS), mutations in *SF3B1* (and other spliceosome genes) can also lead to other differential splicing of other alternative patterns. As

described in Figure 1.2, these include alternative 5' splice site events (A5SS), skipping exon (SE) events, retention of intron (RI) events and mutually exclusive exon (MXE) events. Despite modeling studies of *SF3B1* mutations and the outlining of the genetic landscape in MDS, the determination of the downstream effectors of *SF3B1* mutations remains unclear.

Approximately 30% of MDS patients progress to AML resulting in chemoresistant disease and extremely poor outcomes (5-year overall survival of <30%) [39,40]. Molecularly, progression to AML can take place in different ways [23]. Mutations leading to leukemic transformation can be either acquired as secondary events (e.g. *EZH2* in *SF3B1*-mutated MDS) or be present when clinical symptoms of MDS appear, but expand and attain dominance at a later stage under selection pressure (such as *RUNX1* or *STAG2*) [23]. Upon MDS progression to AML, blasts accumulate and reach the 20% diagnostic threshold for AML [23].

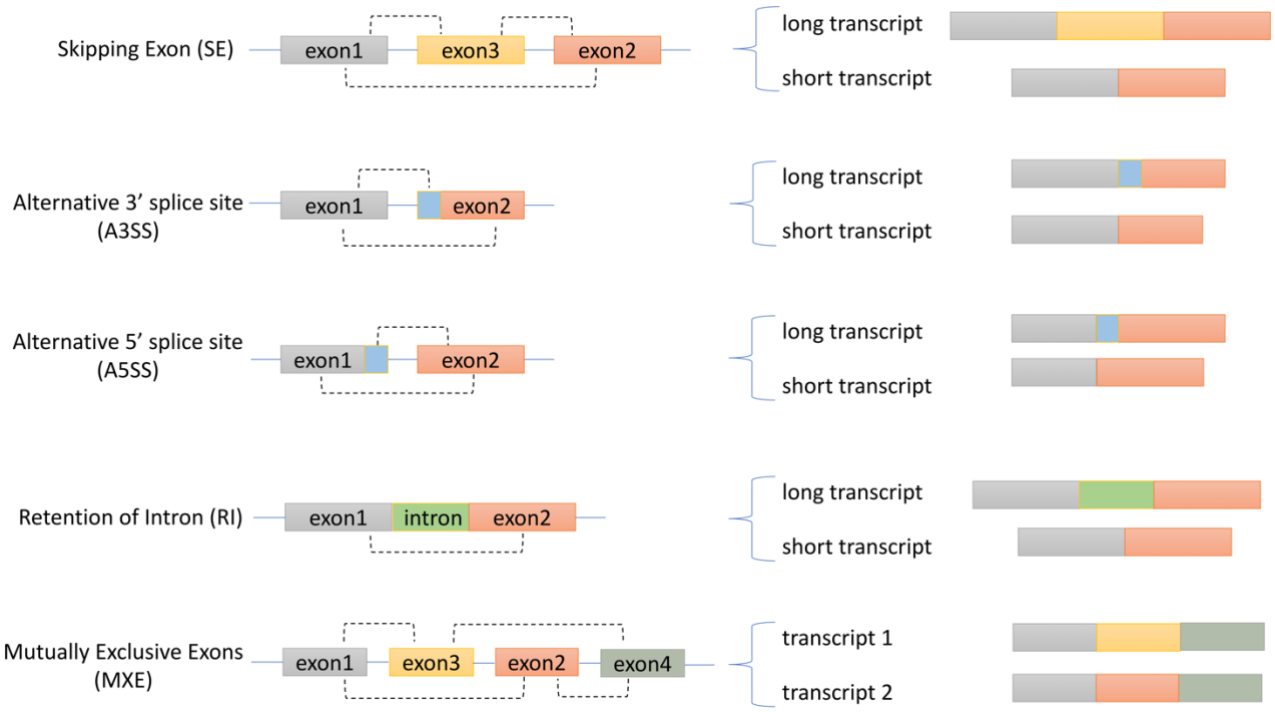


Figure 1.2. Types of alternative splicing events. 1) Skipping Exon (SE): An exon (exon 3 in the figure) may be excluded from the transcript or retained. 2) Alternative 3' splice site (A3SS): An alternative 3' splice junction of exon 2 is used. 3) Alternative 5' splice site (A5SS): An alternative 5' splice junction of exon 1 is used. 4) Retention of intron (RI): The intronic region may or may not be spliced out of the transcript. 5) Mutually Exclusive Exons (MXE): Only one out of two exons (exon 2, exon 3) participates in the transcript.

Chapter 2 of this thesis introduces a multi-stage fusion strategy to integrate distinct transcriptomic readouts from the splicing, transcript and gene level, as well as incorporates analyses of chromatin accessibility data to characterize the functional implications of *SF3B1* mutations in MDS.

1.3.2 Acute Myeloid Leukemia

AML describes a set of aggressive and clonal MNs with rapid onset, acute progression and frequently chemoresistance disease [41–44]. Similar to other MNs, AML pathogenesis is characterized by the presence of a differentiation block that impairs hematopoiesis by preventing progenitor cells from proceeding toward more mature myeloid cell types. This causes the aggregation and expansion of abnormal immature hematopoietic precursor cells called blasts [41–44]. Patients with AML either can be asymptomatic with abnormal complete blood count (CBC) or, in the majority, present with symptoms associated with BM failure (e.g. fever, infections, anemia, bruising, etc) [45,46]. AML has a diverse age range, affecting most often older individuals (the median age of diagnosis is 68 years old) and is associated with poor outcomes. Specifically, disease occurrence rises with age and mortality is higher than 90% when the age of diagnosis is higher than 65 years [45,47].

AML can arise as a de-novo disease, may have a preceding MN such as an MDS preface (secondary AML - sAML) or develop as a consequence of prior therapy (therapy-related AML). AML development is the result of the stepwise acquisition of somatic mutations and cytogenetic aberrations in HSPCs that impede differentiation and promote the proliferation and increase of the blast population [47,48]. In clinical practice, diagnosis of AML requires the presence of blasts at a percentage higher than 20%, usually evaluated morphologically on a bone marrow aspirate. The complete diagnostic profiling relies on immunophenotyping by flow cytometry, which can confirm the existence of excessive blasts and other cell type populations, cytogenetic testing (e.g. FISH or karyotyping) for the identification of chromosomal aberrations and genetic screening for the cataloging of mutations [42,44,46,47,49].

The genomic landscape of the disease includes mutations in more than 80 genes, but only a small fraction of them is frequently mutated (>5% of AML patients) [41]. More than 90% of AML patients are identified with somatic mutations and typically most of them are identified with 2-3 drivers [41,49]. Mutation acquisition is defined by well-characterized and preferred patterns of co-mutations that are ordered in time (Figure 1.3). Early and disease-initiating events usually happen on epigenetic modifiers such as *DNMT3A*, *ASXL1*, *IDH1, IDH2* and *TET2 [41]*. These events are part of the parent clone and the fact that they are rarely identified in isolation, shows that, on their own, are not capable of conferring overt leukemic disease. These events are followed by secondary mutations in genes of the cohesin complex or chromatin modifiers (e.g. *BCOR*, *STAG2*), RNA splicing (e.g. *SRSF2*, *U2AF1*, *SF3B1*) genes or transcription factor genes (e.g. *WT1*, *RUNX1*, *GATA2*) [41]. Late mutations, often associated with disease progression, occur usually in signaling genes such as the receptor tyrosine kinase–RAS pathway genes (Figure 1.3). Other recurrently mutated genes in AML are *NPM1* and *TP53*, the genetic lesions of which typically take place as subclonal events [41].

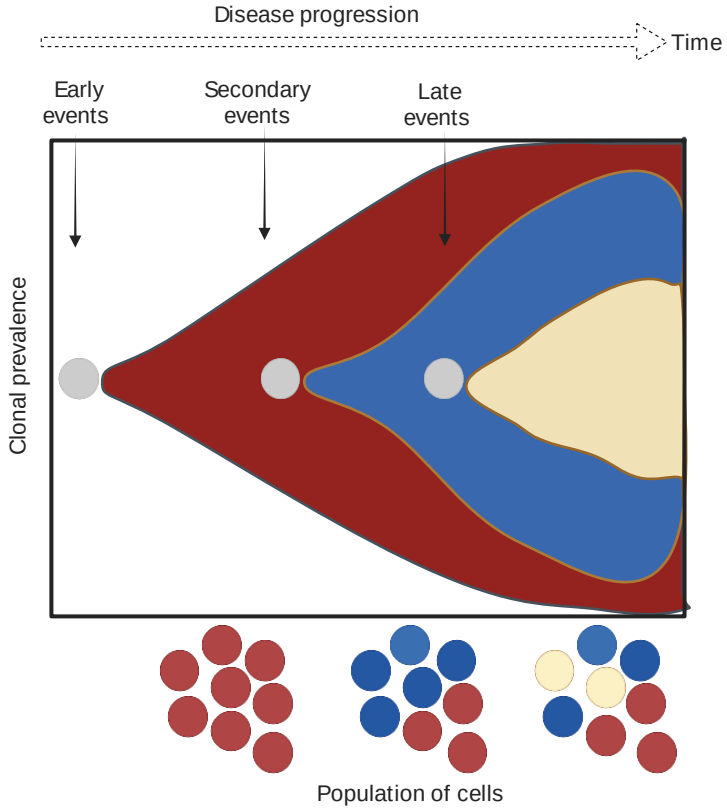


Figure 1.3. Disease progression in AML. Early, disease-initiating events are acquired by myeloid progenitor cells forming the parent clone (red). These events are followed by secondary mutations, usually in chromatin modifiers and spliceosome genes and late mutations in signaling genes[41]. In the course of disease progression, the gradual acquisition of secondary and late events gives rise to subclones (blue, yellow) and creates populations of cells with different genetic landscapes. Drawn using BioRender.

Studying the genomic composition of the disease shows that AML progression follows ordered evolutionary trajectories. The co-occurrence and exclusivity of mutations define heterogeneous and dynamic sets of subclones and outline the need for personalized therapy design. For instance, two of the most frequently mutated genes in AML (15-20% of patients) are *IDH1* (*p.R132*) and *IDH2* (*p.R140* and *p.R172*). Mutations in these genes are early initiating events (clonal) and are frequently identified with co-mutations in *DNMT3A*, *NPM1*, *SRSF2* and *NRAS*. *IDH1/2* mutations cause the elevated production of the oncometabolite 2-hydroxyglutarate (R-2-HG) leading to hypermethylated phenotypes [50–54]. Therapeutic approaches combine IDH inhibitors (e.g. ivosidenib, enasidenib) with chemotherapy or hypomethylating agents and aim to decrease the 2-HG production and relieve the myeloid differentiation block [55–57]. Responses to these therapies are multifactorial and depend at least in part on the presence of resistance-associated mutations such as in the receptor tyrosine kinase–RAS pathway [56,58].

Chapter 3 of this thesis explores single cell transcriptome and genotyping data derived from *IDH1/2* mutated primary AML patient samples, aiming to derive associations between genomic abnormalities and gene expression profiles.

1.4. Omics profiling applications in cancer research

Investigating the role of mutations in disease pathogenesis and evolution relies on the study of primary patient samples as well as on experimental models of disease biology. Examples of such models with applications in myeloid neoplasms include murine models [59,60], immortalized cell lines [61,62] and induced pluripotent stem cells (iPSCs) [63,64]. Murine models can be combined with genetic engineering approaches to introduce mutations as seen in human samples. Immortalized cell lines, such as the leukemic K562, represent primary patient cells that have been engineered to grow in vitro, while iPSCs represent primary patient cells that have been reprogrammed back into an embryonic-like pluripotent state that can be subsequently differentiated into the cell lineage of interest [65,66]. The deployment of omics profiling across cells derived from models of disease biology or primary patient samples, creates an opportunity to study cellular states and characterize putative mechanisms that are directly linked to acquired gene mutations and implicated in disease pathogenesis.

The present thesis leverages omics profiling data from primary patient samples (Chapter 3), as well as patient-derived iPSC models (Chapter 2).

1.4.1. Bulk omics modalities: Genome, transcriptome and epigenome sequencing

The decrease in the costs together with the high throughput and depth of next generation sequencing (NGS) technologies, have enabled bulk data generation at massive scales at the human genome, epigenome, transcriptome, metabolome and proteome level[67,68].

Genomic sequencing applications allow for the detailed documentation of the mutations present in a genome [69]. DNA-sequencing (DNA-seq) applications include 1) targeted sequencing approaches, where select regions of the genome are captured and sequenced 2) whole exome sequencing (WES) approaches that analyze the DNA sequence of the coding part of the genome 3) whole genome sequencing (WGS) approaches that capture the entire genome (coding and non-coding). Whilst targeted and whole exome sequencing approaches capture small mutations and profile copy number abnormalities in select regions of the genome, WGS allows the analyses of all classes of mutations to include copy number abnormalities and genomic rearrangements. Increasingly, primary patient samples are profiled for the genes most commonly mutated in cancer using targeted genome profiling approaches. These deliver information on the genes mutated in each sample as well as the relative clonal representation of each mutation. In the present thesis, data derived from targeted gene profiling of primary patient samples have directly informed the selection of patient samples for disease modeling (i.e. iPSC generation of MDS patient samples with isolated *SF3B1* mutations, Chapter 2) or phenotyping (i.e. selection of AML patient samples with clonal *IDH* mutations, Chapter 3).

The transcriptome represents the full set of RNA transcripts produced in a cell(i.e. mRNAs, rRNAs, tRNAs, miRNAs and other ncRNAs, such as siRNAs, snRNAs, lncRNAs) and depicts the profile of diverse cell types and highly dynamic cellular states [70]. In addition to offering, high throughput and quality measurements of gene expression, RNA sequencing (RNA-seq) also provides multi-faceted information on alternative

splicing (Figure 1.2), transcript usage (relative expression-abundances of transcripts from the same gene, Figure 1.4) and chimeric gene fusions[71]. Quantitative analyses from RNA-seq lead to detailed insights on the regulation of cancer pathways and mechanisms, the evaluation of the tumor response to treatment as well as the identification of biomarkers in the form of novel isoforms and fusion transcripts [69,72].

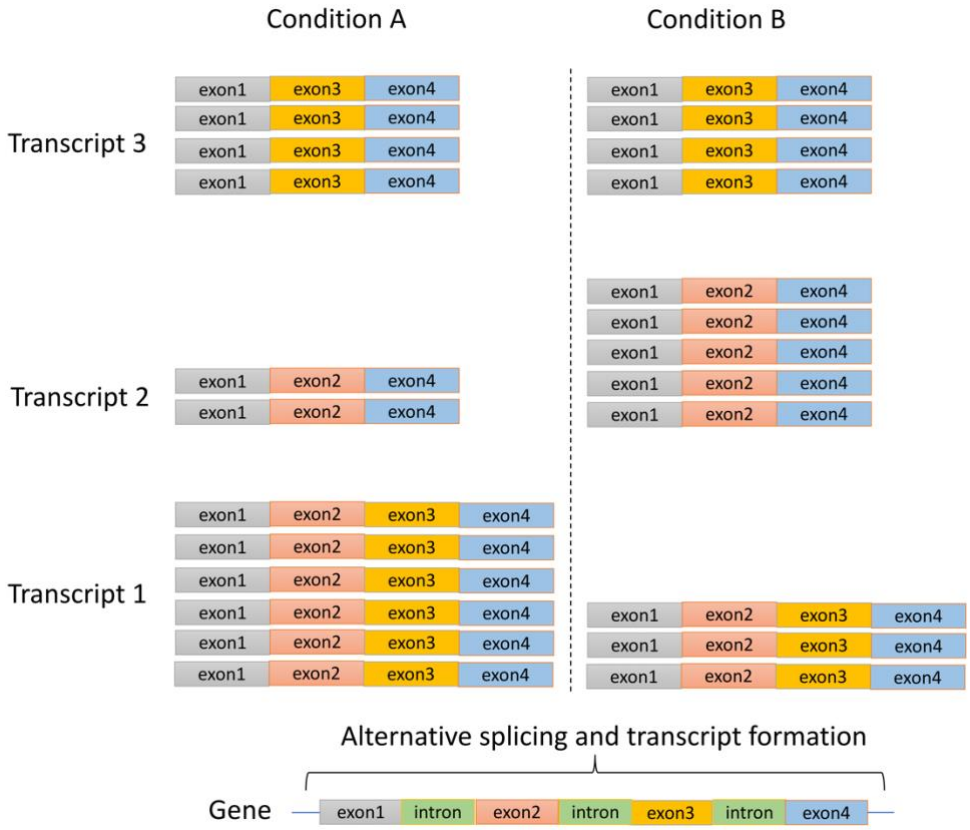


Figure 1.4. Example of transcript usage for a gene across two conditions. Upon the transcription process, the gene is alternatively spliced. This gives rise to 3 different transcripts. The gene has approximately the same overall expression between the 2 conditions, but the proportions of the transcript copies differ. Here transcript 1 has higher usage in condition A, while transcript 2 shows higher usage in condition B.

Epigenomic changes are a diverse set of chemical modifications of DNA nucleotides and histone proteins [73]. They are present genome-wide and have the capacity to confer stable and heritable changes in the functional output of the genome without altering the genomic DNA sequence itself [69,74]. Epigenomic changes include, but are not limited to, DNA methylation (transfer of a methyl group to the C5 position of the cytosine ring of DNA) and post-translational modifications of histones, and alter the chromatin dynamics and accessibility [70,73,74]. Chromatin accessibility or nucleosome positioning along the genome can be profiled using a range of assays such as deoxyribonuclease I (DNase I) hypersensitive site sequencing (DNase-seq), micrococcal nuclease sequencing (MNase-seq) and transposase-accessible chromatin using sequencing (ATAC-seq, Figure 1.5)[69]. The latter is a high-throughput, fast and sensitive technique that can detect genome-wide regions of open chromatin and provide indirect insights on the regulatory processes of gene expression[75].

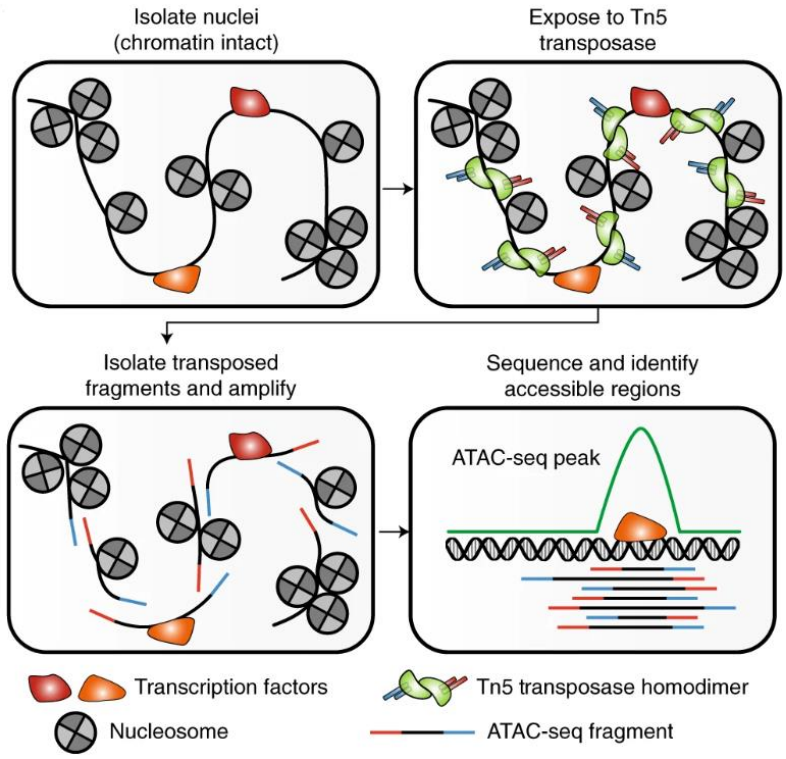


Figure 1.5. Steps of ATAC-seq for chromatin accessibility. Initially, nuclei are isolated from cells retaining their chromatin architecture. Then, the Tn5 transposase cleaves the chromatin regions and the resulting fragments are tagged with adapters. The library fragments are sequenced and ultimately the genomic regions enriched for Tn5 transposition events are emerging as chromatin accessibility (ATAC-seq) peaks. This figure is adapted from Grandi et al. 2022 [73].

In Chapter 2, we harness the breadth of information present in the transcriptomic data (splicing, transcript usage, gene expression) of patient-derived iPSC lines in *SF3B1*-mutated MDS to document the effects of the *SF3B1* mutations in the transcript repertoire. Concurrent analysis of bulk ATAC-seq data from the same iPSC model adds a detailed overview of the chromatin accessibility landscape.

1.4.2. Single cell omics modalities: Genome and transcriptome sequencing

Widespread application of next-generation sequencing approaches have so far been used for the analysis of DNA/RNA extracted from bulk samples which typically capture aggregated information from millions of cells at a time. Although such experiments can be performed at scale and reduced costs generating high quality data, the resulting molecular measurements are averaged across all cells in a sample [69,74] (Figure 1.6). Thus, the study of biological variability among cell subsets such as cells of distinct lineages or clones is not feasible from bulk sequencing approaches. To this end, single cell technologies enable the examination of omics profiles at the single cell level, thereby providing high resolution cell-specific molecular measurements that can be studied in unison as well as in aggregate [15,69]. These data enable the unmasking of the cellular diversity between different populations and the robust exploration of intra-tumor heterogeneity (genetic and clonal)[76] (Figure 1.6).

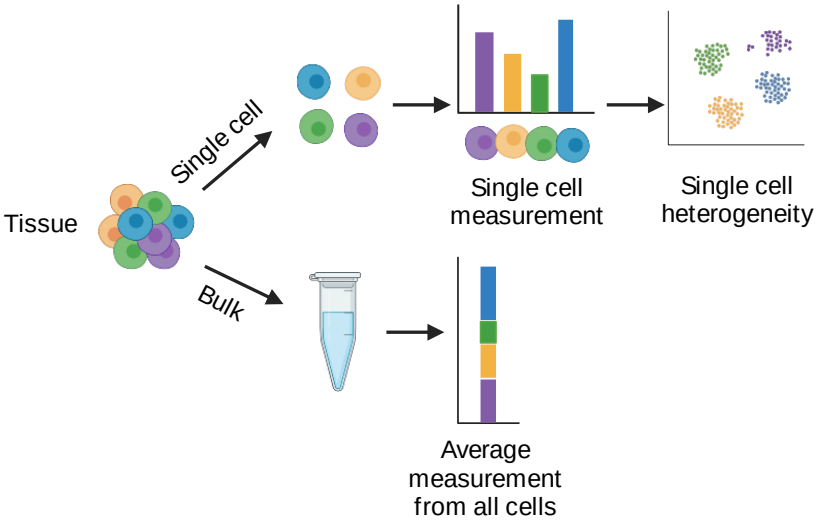


Figure 1.6. Bulk vs Single cell sequencing. Bulk sequencing provides average measurements from all cells of the tissue. On the other hand, single cell sequencing technologies allow measurements for each single cell profile and thus they unmask cellular heterogeneity. Adapted from 10X Genomics and drawn using BioRender.

Current single cell assays profile the genome, transcriptome, epigenome, proteome and metabolome allowing the investigation of the genetic diversity of clones, gene expression dynamics, chromatin accessibility states, transient protein abundances and metabolic changes, respectively[77]. Among all available types of single cell omics data, single cell RNA-sequencing (scRNA-seq) is the most widely used application [78]. Particularly, scRNA-seq data captures the transcriptional states of different cell populations and permits the comparison of gene expression profiles between single cells, leading to the dissection of the transcriptional diversity between distinct clones[76]. As a result of comparing cell identities, single cell transcriptomic analyses uncover cell-to-cell heterogeneity, aid the identification of rare or novel cell types, shed light on previously unknown differentiation trajectories and reveal the dynamics of regulatory networks and pathways[78]. Exploiting this potential of the raw scRNA-seq data has been facilitated by the development of new computational methods or the novel application of existing ones to the single cell field. Specifically, a wide set of methods focuses on reducing the dimensionality of scRNA-seq data, projecting cells onto low dimensional spaces (e.g. t-SNE, UMAP) and grouping cell populations through alternative clustering techniques (e.g. Leiden[79], pcaReduce[80], SC3[81]). Another category of algorithms aim to impute missing gene expression values (e.g. MAGIC[82], scImpute[83]) due to the phenomenon of dropout, i.e the undersampling of mRNA molecules caused by the lack of detection of an expressed gene, especially in the case of lowly expressed genes. Other methods set out to address the inference of regulatory networks (e.g. SCODE[84], SCENIC[85]) while other approaches concentrate on inferring topologies for cellular trajectories describing the temporal evolution of cells (e.g. DPT[86], PAGA[87]).

Single cell DNA-sequencing (scDNA-seq) documents genomic variation at a single cell resolution and is used for the identification of mutations and copy number alterations at the single cell level [78]. However, derivation of single cell genetic data is more challenging than scRNA-seq. The latter exploits the presence of thousands of RNA copies of each transcript, whereas there are only two copies of DNA in each human cell

[88]. Whole genome amplification (WGA) methodologies offer the opportunity to address this challenge by amplifying or generating multiple copies of genomic DNA. However, at times, WGA techniques might not be able to detect a specific allele within an extensive genomic region and thus struggle to maintain a consistent sequencing depth along the genome[88]. For this reason, bioinformatics tools designed to identify single nucleotide variants (SNVs) attempt to take allele coverage biases and amplification artifacts into account (e.g. SCcaller[89], Monovar[90], LiRA[91]).

1.5. Integration of omics data

It is evident that a variety of omics data types, both at the bulk and single cell level, permit the characterization of molecular properties with unparalleled scale and precision, thereby offering a comprehensive understanding of tumor behavior[69]. However, each omic technology is targeted to a specific molecular type (e.g. genome, transcriptome, epigenome, proteome, metabolome) and thus, the spectrum of insights that can be drawn from single omics studies is limited to the biological scope of the measured biomolecules[68]. Therefore, such approaches do not have the means to investigate the intricate complexities across biological landscapes and are not sufficient enough to establish relationships among the various molecular layers or decode the dependencies between features from different modalities[71,92]. Despite diving into a higher resolution space assisting in deciphering tumor heterogeneity, even single cell datasets of a single data type may depict a modality-specific view of the cell state and may not have the power to provide a detailed understanding of the function of cellular components and their interactions[77].

1.5.1. Multi-view mining with bulk omics: Background & Benefits

Hence, beginning from the bulk level, bringing together knowledge from multiple sequencing data modalities or analyzing in unison the full spectrum of information offered from a single sequencing technology (e.g. RNA-seq), results at a more complete perspective of the underlying disease biology[93]. Considering each omic modality or each set of measurements from a specific data type (e.g. gene expression, transcript usage, alternative splicing from RNA-seq data) as a separate data view, then the integration of all these data views together utilizes their complementary nature and leads to synergistic conclusions influenced from various biological aspects[94]. Such multi-view studies allow the analysis of a range of high-dimensional measurements at multiple levels and scales and thus, can comprise the backbone of data-driven translational research[71]. Additionally, using various views of data together, can help overcome any uncertainty generated by any missing or unreliable information in any single view[95]. Overall, bulk integrative approaches are empowered to examine the flow of molecular information between the multiple data views[68], display a holistic picture of molecular systems[96], processes and mechanisms at the tissue level and elucidate complex cancer phenotypes [67].

In particular, multi-view data integration within the setting of a single bulk sequencing modality such as RNA-seq, can be interpreted as the systematic assembly of multiple types of measurements (views) extracted from the same source[97]. In this case, these views depict distinct profiles or levels of the same data (e.g. alternative splicing, transcript usage, gene expression) and jointly represent the entire spectrum of

knowledge provided by the respective omics type (transcriptomics)[71,98]. Integrating all branches within the same omics modality in a multi-level fashion, interrogates molecular properties from different levels and traverses the flow[71] of information from one level to the other. This process uncovers the underpinnings of biological processes[74], and ultimately links characteristics of biomolecules across multiple stages, aiding the formation of a comprehensive representation of the underlying landscape. Therefore, combining supporting evidence from these different levels can explain previously reported observations from single-view analyses and also lead to the identification of novel biomarkers. For example, Ha et al. 2021[97] jointly consider gene expression, splicing and polyadenylation patterns from RNA-seq data to examine the relative contribution of multiple transcriptomic regulatory layers in the specification of neuronal identities. Their results highlight the significance of coordinating multiple aspects of the same transcriptomic data in the framework of defining the temporally and spatially distinct neuronal subpopulations.

Alternatively, in the setting where data views correspond to unique bulk sequencing modalities, multi-view data integration typically unites the profiles of different types of biomolecules originating from the same samples. Multi-omics approaches with bulk sequencing data focus on documenting the interactions between biological layers[99] as well as describing the regulatory mechanisms that could possibly explain complex phenotypic traits or the behavior of molecular systems[93]. Bringing together complementary omics-views enhances a more thorough disease profiling and, based on the scope of the study, can also lead to an improved disease subtyping and patient stratification[93,100], the identification of multifaceted biomarkers for cancer diagnostics[93,100], the improved prediction of clinical outcomes[101] and a more complete understanding of the responses to therapy[100]. Collating views of the genome, transcriptome, and epigenome allows a more precise characterization of cancer biology and shows how gene expression patterns are related to chromatin accessibility topologies and histone modifications within different genetic contexts. For instance, Xiang et al. 2020[102] merged epigenetic features and transcriptomes to produce a detailed picture of the regulatory landscape of differentiating hematopoietic cell types in mice. On a similar note, Chen et al. 2022[103] utilize collectively RNA-seq, ATAC-seq and ChIP-seq (chromatin immunoprecipitation assays used for the identification of genome-wide DNA binding sites for transcription factors and other proteins) data to reflect the role of JUNB in the human hematopoietic fate induction.

1.5.2. Multi-view mining with single cell omics: Background & Benefits

Integration of omics modalities at a single cell level can be either matched, i.e. different omics data are captured from the same cell or unmatched, i.e. different omics data are captured from different samples[104]. Focusing on the former case, matched data integration necessitates experimental protocols that process more than one type of biomolecules. The joint profiling of genotypes and transcriptomes occurs either through low throughput techniques such as scSIDR-seq[105] or medium throughput ones such as G&T-seq[106] or higher throughput ones such as TARGET-seq[107]. Despite the concurrent ascertainment of DNA and RNA from the same cell, the cost, technical challenges and labor-heavy requirements of these protocols[108], pose significant considerations for their adoption at scale. An alternative approach for the integration of genotype information to transcriptomic data has been to call somatic variants and identify copy number variations fromscRNA-seq data alone, using bioinformatics tools such as SComatic[109],

Monopogen[110] and InferCNV[111], HoneyBADGER[76] respectively. Alternatively, Nam et al. 2019[112] presented an experimental framework, called GoT (genotyping of transcriptomes), that is able to link select genotypes to the transcriptional profiling of thousands of single cells using scRNA-seq data. Using a modification of the 10X scRNA-seq pipeline and the development of allele specific genotyping probes, GoT enables targeted genotyping of scRNA-seq libraries. This enables the derivation of somatic genotypes for a set of a priori known variants across thousands of cells. However, genotyping efficiencies may differ significantly across alleles and are dependent on the expression levels of the gene of interest and the proximity of the desired allele to the 5' or 3' of the transcript[112]. Chapter 3 leverages genotypic profiles derived from the application of GoT on scRNA-seq data from *IDH1/2*-mutated AML primary patient samples

Single cell multi-omics approaches, by profiling genetic abnormalities and the transcriptome, have the capacity to reveal genotype-phenotype associations[78]. For example, genotype-phenotype associations can be established as the links between specific genetic variations and their downstream impact on the expression of disease driver genes[78]. Considering intra-tumor heterogeneity, the joint analysis of single cell transcriptomic data and genotypes enlightens the correspondence between cellular states or transcriptionally unique cell populations with genetic subclones[76,113]. This integrative path leads to the functional characterization of specific mutations or cytogenetic aberrations and the deciphering of their consequences on distinct phenotypes. For instance, Macaulay et al. 2015[106], by applying G&T, discovered a subset of HCC38-BL (B lymphoblastoid cell line) cells with trisomy of chromosome 11. This subpopulation exhibited elevated expression in the genes of chromosome 11 compared to the diploid cells. Additionally, Rodriguez-Meira et al. 2019[114], by using TARGET-seq, identified abnormal expression of oncogenes (like *MYCN*, *TP53*) as well as of Wnt signaling and interferon-related genes in $JAK2^{V617F}$-mutated HSPCs. Moreover, examining such approaches within the inherent complexity of AML, makes the coupling of single cell transcriptomic with genomic data a necessary step toward understanding the connection between genetic and transcriptional heterogeneity. In this direction, Petti et al. 2019[113] examined the cell representation of identified AML gene expression clusters at the phenotypic and mutational level in an unsupervised setting while van Galen et al. 2019[115] similarly combined single cell genotypes with gene expression measurements to correlate cell type compositions with genetic lesions.

Chapter 3 leverages transcriptomic and genomic information from the same cells and captures links between gene expression profiles and genetic abnormalities in *IDH1/2* mutated AML.

1.5.3. Multi-view integration strategies

From a computational perspective, integrating multiple views of data (representations or sets of features derived from the measured biomolecules either within or across modalities) consists of two overlapping procedures: data fusion and data interconnection[116] (Figure 1.7). Data fusion describes the process of extracting and combining complementary contextual information from multiple data to improve decision making and the quality of relevant outcomes (e.g. for prediction, classification, or clustering tasks). Data interconnection denotes the action of unmasking the information and associations shared between data views[116] (Figure 1.7).

Data fusion strategies

Multi-stage fusion

Methods for data fusion can be broadly divided into two categories: the multi-stage and the meta-dimensional[95,96] (Figure 1.7, Table 1.1). In the multi-stage fusion, information is integrated in a stepwise or hierarchical mode. In particular, multi-stage approaches handle data views as separate entities and build frameworks that sequentially or hierarchically enrich the emerging signals with additional layers of information[95,96]. This strategy, as the name suggests, divides the analysis to distinct steps and may involve some degree of independent processing of each view before investigating inherent correlations. Results derived from the analysis of each view may serve as anchors between the different steps and can facilitate the framework to 1) establish inter-layer associations 2) relate these associations with a phenotype in a supervised setting or produce new conclusions in an unsupervised manner.

Depictive applications of this multi-stage logic in a biological or molecular context can be seen in bulk omic approaches such as genomic variation analyses, domain knowledge guided analyses and allele-specific expression (ASE) analyses[95] (Table 1.1). The genomic variation analyses attempt to associate variations in the DNA with another data view (usually gene expression, DNA methylation or protein levels) and then, relate both of them with a phenotype of interest. This can be achieved either through likelihood-based causal inference approaches[117] or, most commonly, through a three-stage technique[95]. In the setting of the latter, initially, genomic variation analyses identify single-nucleotide polymorphisms (SNPs) associated with the phenotype of interest. Second, the statistically significant SNPs from step 1 are tested for association with the second omic view (e.g. gene expression). In the final step, the data from the second omic view are correlated with the phenotype of interest. For instance, efforts that adopted this approach combined genome-wide SNPs with baseline gene expression levels of HapMap lymphoblastoid cell lines (LCLs) to identify associations with the IC50 drug cytotoxicity measurements[118–120]. Domain knowledge guided analyses also adopt a multi-level logic, using at the same time derived insights either from known biological mechanisms and processes or from documented functional and pathway information in resources such as ENCODE[121] and KEGG[122]. Compared to the genomic variation analyses, the domain knowledge guided ones involve an additional step of annotating genetic variants and selectively advancing only a subset of them to the subsequent stages of the analysis[95]. Lastly, ASE first assesses whether the maternal or paternal allele is preferentially expressed and then relates this allele to variations of *cis*-regulatory elements and epigenetic modifications[95]. ASE as well as its extensions such as allele-specific transcript structure, have been employed to discover functional variation[123] and protein-DNA[124] interactions in humans.

Chapter 2 uses principles from the multi-stage fusion strategy to integrate different levels of deregulation of the transcriptome (splicing, transcript usage, gene expression) in *SF3B1*-mutated MDS.

Beyond the multi-stage fusion strategy utilized in this thesis, there has been active development of a plethora of meta-dimensional (concatenation-based, transformation-based and late) fusion approaches. In the section below, we highlight some notable and relevant innovations of this fast growing field.

Meta-dimensional fusion: Concatenation-based & late strategies

Meta-dimensional fusion strategies merge diverse data views simultaneously and can be divided into three categories: concatenation-based (early) fusion, transformation-based (intermediate) fusion and late fusion[95,96] (Figure 1.7, Table 1.1). Concatenation-based fusion methods are usually deployed in bulk sequencing settings and as the name suggests, combine all data views at the original, hand crafted or lower-dimensional space into one common representation (e.g. matrix) before constructing a model[69,92,95,96,116]. This representation can be used as a single entity and serves as input for any downstream modeling (Figure 1.7). This straightforward strategy, by simply concatenating data views at an early stage, allows the use of a variety of downstream models for mining the data and has the advantage of taking feature interdependencies and interactions into account[92]. However, the resulting representation might suffer from the 'curse of dimensionality' due to the size of the final feature space[69]. For instance, Fridley et al. 2012[125] fused data from SNPs and gene expression into a single matrix and applied Bayesian modeling to predict drug cytotoxicity. Mankoo et al. 2011[126], after computing Spearman rank correlations among different data types, applied a multivariate Cox Lasso model on a merged representation of CNVs, methylation and gene expression data to predict survival in ovarian cancer.

Contrary to concatenation fusion methods, late fusion methods conduct independent analysis for each data view and then consolidate the individual results[69,92,95,96,116]. In particular, first, unimodal models are constructed using single views and in a second step, a final model is built from them following a reasoning strategy (Figure 1.7). Popular choices for the development of the final model are, among others, majority voting, multiple kernel learning as well as ensemble approaches. The limitation of the late fusion strategy is its inability to account for the complementarity and the interdependencies of the different data views[95]. On the other hand, the fact that the integration is applied on the independently derived outcomes of unimodal models, makes the late fusion strategy agnostic to the method used to derive each individual outcome. Thus, it provides the flexibility of changing modeling approaches without affecting the architecture and the reasoning of the fusion method. In a bulk sequencing context, Tao et al. 2019[127] deployed multiple kernel learning on CNV, transcriptomic and methylation data for subtype prediction in breast cancer. Additionally, for cancer subtyping tasks, Aure et al. 2017[128] applied COCA (cluster-of-clusters analysis)[129] to fuse clustering assignments produced on multiple omic levels (e.g. protein, miRNA, metabolic profiles). At the single cell level, late fusion approaches can be applied for the integration of clustering results from different algorithms on the same data modality (e.g. scRNA-seq). For example, Huh et al. 2020[130] fused clustering annotations from multiple clustering methods via mixture model ensemble solving a maximum likelihood problem, and provided a consensus clustering of human peripheral blood mononuclear cells (PBMCs).

Meta-dimensional fusion: transformation-based strategies

Transformation-based fusion approaches jointly analyze all data views within a modeling framework, commonly by mapping or transforming the initial data to intermediate representations (e.g. graphs, kernel matrices, etc)[69,92,95,96,116] (Figure 1.7). In such approaches, the fact that modeling occurs on the transformed space, facilitates the integration of measurements across different scales and types (discrete,

continuous)[95]. At the same time, the transformation of the data views to intermediate representations does not mask data specific properties, but might render the recognition of interaction effects between the single views more challenging[69]. The categorization of the transformation-based methods is not definite and it can depend on technical aspects (e.g. deep learning methods, graph-based methods, etc), the type of the transformed representation (e.g. matrices, networks, etc), as well as, the biological problem tackled (e.g. disease subtype identification, patient classification, prediction of clinical outcomes, biomarker identification etc). Here, in an effort to put both bulk and single cell methods under the same perspective and at the same time account for the technical characteristics of the methods used, we categorize transformation-based fusion approaches into three non-exclusive groups: 1. similarity-based methods; 2. dimension reduction-based methods; and, 3. statistical modeling-based methods.

Similarity-based approaches typically learn sample affinities commonly depicted as similarity matrices or graphs for each data view and then merge these inter-sample similarities in a unified context. In particular, similarity network fusion SNF[131], suited for bulk multi-omics efforts, models pairwise patient similarities from each view in view-specific graph structures called similarity networks. Then, SNF integrates these graphs using an iterative fusion procedure that eventually produces a single network, representative of the full spectrum of the underlying data. Wang et al. 2021[132], within a supervised setting for patient classification, handled the view-specific similarity networks as graph convolutional networks (GCNs). The predictions of these GCNs were passed to a view correlation discovery network that produced the final labels. In an alternative approach, Ramazzoti et al. 2018[133] proposed a cancer subtyping method, CIMLR (Cancer Integration via Multikernel LeaRning), which constructs a unified similarity matrix for downstream clustering, after combining multiple gaussian kernels per view. In single cell omics, Hao et al 2021[134] generated a weighted nearest neighbor graph which, for each cell, depicts its most similar ones based on a weighted combination of the different modalities (e.g. protein and RNA). This graph can be used for downstream analyses such as low dimensional projections and clustering of PBMCs. Alternatively, Singh et al 2021[135], first chose one modality as the primary one, then determined modality specific cell similarities and finally transformed the primary modality such that it has maximum level of agreement with the rest. In another approach, CiteFuse[136] computed the cell-to-cell similarity matrices of matched proteomic and mRNA data separately and then applied SNF to merge them. Then, graph-based clustering was performed on this merged similarity matrix.

Dimension-based reduction strategies typically deploy matrix factorization (MF), canonical correlation analysis (CCA) or deep learning (DL) techniques and focus on projecting the different data-views to a joint latent space. In particular, joint factorization decomposes the observed data matrices into sets of low dimensional latent factors able to capture inter-view dependencies. Lock et al 2013[137], in the context of bulk multi-omics, proposed JIVE (Joint and Individual Variation Explained), a MF framework that decomposes data variation into a component representative of the joint biological variation across views and into another component that is specific for each view. Argelaguet et al 2018[138], using MF, introduced MOFA (multi-omics factor analysis), a framework that operates both in bulk and single cell data and infers latent factors that capture the underlying sources of variation across either complete or partial data views. MOFA was applied on a chronic lymphocytic leukemia cohort to identify axes of disease heterogeneity. Welch et al

2019[139], developed LIGER (linked inference of genomic experimental relationships), a method that uses integrative non-negative MF[140] on single cell multi-omics to represent cell profiles as a mixture of view-specific and shared factors. LIGER can be used for cell type identification from scRNA-seq and scATAC-seq.

CCA, as a subspace learning approach that aims to identify pairs of projections for different views such that correlations between them are maximized, can be also used as a backbone for multi-omic fusion efforts. For example, in a bulk setting, CCA[141,142] provides insights on inherent structures by spotting correlated patterns across different omics types (such as finding genomic regions with CNVs correlated with various expression levels). In a single cell context, Stuart et al 2019[143] use CCA to jointly reduce the dimensionality of two single cell datasets and then search for mutual nearest neighbors in the shared low dimensional space. This method can be applied for the integration of multiple single cell measurements from different scRNA-seq samples and technologies. In another approach, MAESTRO[144] projects cells with matched transcriptomic and chromatin accessibility data into a unified low-dimensional space by performing CCA between gene expression from scRNA-seq and regulatory potential from scATAC-seq.

Apart from MF and CCA, deep learning architectures and specifically autoencoders are also commonly deployed for transformation-based fusion approaches. They have the ability to learn unified low dimensional embeddings across views and capture nonlinear dependencies at the same time. Even though the deployment of autoencoders can be also seen in bulk multi-omic integration[145], the fact that single cell data sets typically comprise thousands of cells, offers more suitable conditions for optimal training and exploitation of their computing capacities. Within a single cell setting, Gayoso et al. 2021[146] introduced totalVI (Total Variational Inference), a probabilistic framework that learns representations of cell profiles based on the joint low-dimensional embeddings of RNA and protein data emerging from a variational autoencoder architecture. These representations can be used for visualization purposes and downstream tasks such as cell clustering. Additionally, Lin et al 2022[147] proposed a deep autoencoder model, named scMDC (single cell Multimodal Deep Clustering), that also jointly forms latent features of encoded embeddings for the clustering of matched single cell multi-omics views (e.g. from PBMCs).

Statistical modeling-based methods allow the joint probabilistic modeling of multi-omics data under a Bayesian framework. Lock & Dunson et al. 2013[148] introduced Bayesian consensus clustering (BCC), a Bayesian framework that utilizes Dirichlet mixture models to produce view-specific but dependent clusters that adhere loosely to an overall consensus clustering. Application of BCC on bulk transcriptomic, methylation and proteomic data from TCGA[8] (The Cancer Genome Atlas) resulted in breast cancer subtypes with specific clinical features. In a similar approach, Kirk et al. 2012[149] describe a Bayesian method, named Multiple Dataset Integration (MDI), which does not assume a common clustering structure, but instead defines view-specific clusters while also modeling the pairwise dependencies between them. Alternatively, in a non-inherently Bayesian setting, other efforts determine a single joint clustering by finding the structure that maximizes a joint likelihood. This approach is followed by Kormaksson et al. 2012[150] upon fusing gene expression and DNA methylation data, while the iCluster[151] method initially fits a Gaussian latent factor model to the joint likelihood and then applies K-means on the factor scores to extract the cluster assignments. iCluster has been applied for breast and lung cancer subtyping based on copy number and gene expression data. At the single cell level, contrary to the similarity-based and autoencoder-based methods,

the statistical modeling-based ones have not been widely deployed for multi-view fusion. However, Wang et al. 2020[152] developed BREM-SC, a hierarchical Bayesian mixture framework that adopts Dirichlet multinomial distributions to model the expression levels of genes and surface proteins from matched single cell data. Wang et al. also introduce in their approach cell-specific random effects to model the correlation between these two data views and apply BREM-SC for the clustering of publicly available single cell data from PBMCs.

Table 1.1. Literature examples of multi-view omics integration across fusion strategies.

Fusion strategy	**Method**	**Sequencing type**	**Omics types used**	**Scope**	**Reference**
Multi-stage	QTDT*, general linear models	Bulk	Genomics, transcriptomics, cytotoxicity assays	Genomic variation, biomarker identification	[118–120]
Multi-stage	Likelihood-based causal inference	Bulk	Genomics, transcriptomics	Genomic variation, biomarker identification	[117]
Multi-stage	Maximal concordance	Bulk	Genomics, transcriptomics	Allele-specific expression, functional variation identification	[123]
Multi-stage	Significance analysis of microarrays	Bulk	Genomics, epigenomics	Allele specific protein-DNA interactions	[124]
Meta-dimensional (concatenation-based)	Multivariate Cox Lasso	Bulk	Genomics, transcriptomics, epigenomics	Survival prediction	[126]
Meta-dimensional (concatenation-based)	Bayesian modeling	Bulk	Genomics, transcriptomics	Phenotype prediction (drug cytotoxicity)	[125]
Meta-dimensional (late)	Multiple Kernel Learning	Bulk	Genomics, transcriptomics, epigenomics	Cancer subtyping	[127]
Meta-dimensional (late)	COCA*	Bulk	Genomics, transcriptomics, metabolomics, proteomics	Cancer subtyping	[128]
Meta-dimensional (late)	Mixture model ensemble	Single cell	Transcriptomics	Clustering	[130]
Meta-dimensional (transformation-based)	Similarity Network Fusion	Bulk	Transcriptomics, epigenomics	Cancer subtyping, survival prediction	[131]

Meta-dimensional (transformation-based)	Graph Convolutional Networks	Bulk	Transcriptomics, epigenomics	Patient classification, cancer subtyping, biomarker identification	[132]
Meta-dimensional (transformation-based)	Multiple Kernel Learning	Bulk	Genomics, transcriptomics, epigenomics	Cancer subtyping, survival prediction	[133]
Meta-dimensional (transformation-based)	Weighted Nearest Neighbor graphs	Single cell	Transcriptomics, proteomics, epigenomics	Cell state identification, clustering	[134]
Meta-dimensional (transformation-based)	Metric learning	Single cell	Transcriptomics, epigenomics	Modality alignment, cell type inference	[135]
Meta-dimensional (transformation-based)	Similarity Network Fusion	Single cell	Transcriptomics, proteomics	Clustering	[136]
Meta-dimensional (transformation-based)	Matrix Factorization	Bulk	Transcriptomics	Cancer subtype characterization	[137]
Meta-dimensional (transformation-based)	Matrix Factorization	Bulk & single cell	Genomics, transcriptomics, epigenomics	Biomarker identification	[138]
Meta-dimensional (transformation-based)	Matrix Factorization	Single cell	Transcriptomics, epigenomics	Cell type identification, clustering	[139]
Meta-dimensional (transformation-based)	Canonical Correlation Analysis	Bulk	Genomics, transcriptomics	Survival prediction, cancer subtyping	[141]
Meta-dimensional (transformation-based)	Canonical Correlation Analysis	Single cell	Transcriptomics, epigenomics, proteomics	Batch-effect correction, cell state identification, clustering	[143]
Meta-dimensional (transformation-based)	Canonical Correlation Analysis	Single cell	Transcriptomics, epigenomics	Clustering, cell type annotation	[144]
Meta-dimensional (transformation-based)	Autoencoders	Bulk	Transcriptomics	Clustering, cancer subtyping, molecular characterization	[145]
Meta-dimensional (transformation-based)	Autoencoders	Single cell	Transcriptomics, proteomics	Batch-effect correction, biomarker identification	[146]
Meta-dimensional (transformation-based)	Autoencoders	Single cell	Transcriptomics, epigenomics, proteomics	Batch-effect correction, clustering	[147]

Meta-dimensional (transformation-based)	Bayesian modeling	Bulk	Transcriptomics, epigenomics, proteomics	Clustering, cancer subtyping	[148]
Meta-dimensional (transformation-based)	Bayesian modeling	Bulk	Transcriptomics, epigenomics, proteomics	Clustering, molecular characterization	[149]
Meta-dimensional (transformation-based)	Mixture models and likelihood estimation	Bulk	Transcriptomics, epigenomics	Clustering, cancer subtyping	[150]
Meta-dimensional (transformation-based)	Latent variable modeling and likelihood estimation	Bulk	Genomics, transcriptomics	Clustering, cancer subtyping	[151]
Meta-dimensional (transformation-based)	Bayesian modeling	Single cell	Transcriptomics, proteomics	Clustering	[152]

* QTDT: quantitative transmission-disequilibrium test

* COCA: cluster-of-clusters analysis

Data interconnection

The goal of multi-view data interconnection is to establish links and relationships between pairs of data views, as well as, examine how specific patterns visible in one view are represented in the other[116] (Figure 1.7). An important principle of data interconnection is the flow of information from one data view to the other. A classic paradigm of this flow, depictive of disease etiology, is that the downstream consequences of oncogenic variations can be at least in part ascertained through gene expression phenotyping. A common way of connecting gene expression phenotypic profiles with genetic abnormalities at the cell level, is to overlay cell-level genetic annotations (e.g. using genomic profiling to derive mutation information) onto the low dimensional representation of the same cells from matched scRNA-seq. This provides the opportunity to associate gene expression clusters with the malignant or WT status of a sample or cell[115] or the existence of CNVs[113] as well as the mutational status of specific genes[112]. On the contrary, in an unmatched setting, Campbell et al. 2019[153] developed Clonealign whereby each cell's gene expression profile is assigned to its clone-of-origin by integrating independently sampled scRNA-seq (expression) and (scDNA-seq) copy number data. Clonealign is based on a Bayesian latent variable model that maps the copy number of a gene to its expression value by introducing a copy number dosage effect on the gene expression.

In Chapter 3 of the thesis, we deploy deep learning as a means of interconnecting single cell genotypic and expression profiles.

Furthermore, data interconnection is also adopted in the context of exploring the associations between bulk RNA and ATAC-seq data. Such associations are drawn using either the sample or the gene as a point of connection. In the sample-wise manner, chromatin accessibility at regulatory elements is correlated to RNA abundances, while in gene-wise correlations fold changes of differential accessibility and expression analyses are compared[154]. In particular, Sanghi et al. 2021[154] show the density of the sample-wise correlation

between gene accessibility and gene expression from primary and metastatic thyroid samples. In a gene-wise specific context, Wang et al. 2021[155] used diamond plots to present the accessibility fold change of chromatin peaks together with the expression change of the genes associated with them, within different AML genetic subgroups using human iPSCs. Additionally, using the gene as a central anchor allows one to observe the expression of transcription factors matching accessibility motifs[156].

The interconnection between data views also includes combinations of omics with other data modalities. Particularly, the generation of diverse datasets in scale, coupled with the computational efficiency of deep learning models, have enabled links between omic profiles and modalities such as radiology scans, chemical compounds or tissue morphology in pathology slides [116]. For instance, gene expression data have been associated with whole slide (pathology) images (WSIs) and chemical structures. Schmauch et al. 2020[157] used thousands of matched hematoxylin and eosin (H&E)-stained slides and RNA-seq samples across 28 cancer types from TCGA to train HE2RNA, a deep learning model that predicts RNA-seq profiles from histology images without expert annotations. Fotis et al. 2020[158] used Siamese graph convolutions to associate chemical compounds with their affected biological processes inferred from matched gene expression data. In the context of establishing genotype-phenotype associations, Coudray et al. 2018[159] showed that mutations in *STK11*, *EGFR*, *FAT1*, *SETBP1*, *KRAS* and *TP53* can be identified directly from H&E WSIs in lung cancer using the inception v3[160] convolutional neural network (CNN). In MDS, Brück et al 2021[161], connected BM histopathology images to genetic alterations indirectly, through deep CNN features, representative of the tissue morphology. Other studies, in similar settings, aim to predict such mutational statuses in liver[162], bladder[163], colorectal[164] and thyroid[165] cancer, whereas other approaches [166,167] operate on a pan-cancer context. In radiology, Ha et al 2017[168] associated features from breast mammography to the mutational status of *BRCA1/2*, Wang et al 2019[169] trained a CNN using thousands of CT (computed tomography) scans to predict *EGFR* mutations in lung adenocarcinoma, while He et al. 2020[170] deployed a ResNet architecture to predict noninvasive *KRAS* mutations through CT scans in colorectal cancer.

We note that the separation between data interconnection and fusion approaches is not exclusive as both processes inherently follow the idea of leveraging shared or complementary patterns from multi-view profiles. The fact that data fusion techniques might exploit hidden data interconnections as part of their integration strategy, might create an overlap between the two approaches. Concluding, we underline that the choice between the two approaches, as well as, the subsequent selection of computational techniques depends on the task in question, data prevalence, experimental setting as well as computing considerations.

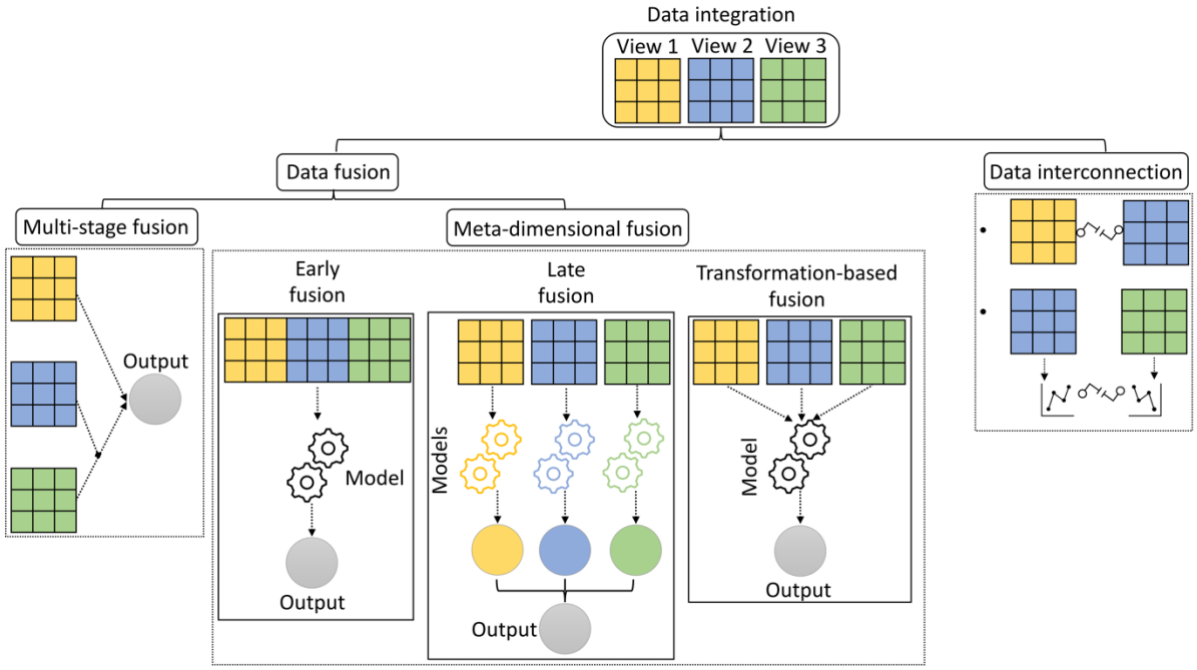


Figure 1.7. Map of multi-view integration in the setting of omics data. Integrating multiple views can be separated into data fusion and data interconnection. The latter examines the cross-talk between pairs of views either directly or indirectly while the former (data fusion) combines complementary contextual information to improve an output. Methods for data fusion can be divided into the multi-stage fusion ones, which merge information from each view in a stepwise or hierarchical fashion, and the meta-dimensional fusion ones which can be further categorized into early (concatenation-based), late and transformation-based approaches. Early fusion techniques first concatenate all views into a single entity and then apply a model on this entity, late fusion techniques perform uni-view analyses and then merge the view-specific outputs whereas transformation-based ones jointly analyze all data views.

1.6. Thesis outline

This thesis adapts concepts from data integration to formally mine omics data derived from primary patient samples or experimental models of MDS and AML. We deploy principles from multi-view data fusion and interconnection to develop analytical frameworks within and between different omics modalities and sequencing techniques, aiming to extract biologically relevant relationships amongst multi-faceted molecular signals. In the context of *SF3B1*-mutated MDS (Chapter 2) and *IDH1/2*-mutated AML (Chapter 3), the presented analyses and frameworks establish links between diverse data views and identify patterns inherent in the data, setting out to elucidate omics landscapes, identify molecular characteristics and assist the study of phenotypes at a genetic level.

Chapter 2 investigates the transcriptional repertoire and chromatin profile of *SF3B1*-mutated MDS, leveraging bulk RNA and ATAC-seq data from patient-derived genetically matched normal and *SF3B1*-mutated iPSC lines. Within this context, we introduce a multi-stage fusion framework which brings together data views from different layers of transcriptome sequencing and results in a detailed overview of the transcriptomic repertoire of *SF3B1*-mutated MDS. Particularly, with a domain knowledge approach, we

merge signals from splicing, transcript usage and gene expression and we derive a splicing signature of 59 splicing events linked to 34 genes, which associates with the *SF3B1* mutational status of primary MDS patient cells. Additional unimodal chromatin accessibility analysis from the ATAC-seq data, showed increased priming of *SF3B1* HSPCs toward the megakaryocyte- erythroid lineage, as well as the enrichment of motifs from the TEA (TEAD) domain in accessible regions linked to genes with upregulated expression. Overall, chapter 2, applies a multi-stage fusion approach on transcriptomic data views to prioritize mis-spliced gene targets, and concurrently provides a formal overview of the *SF3B1*-mutated chromatin landscape and nominates transcriptional programs with putative roles in MDS disease biology.

Inspired by past studies that aim to predict genotypic abnormalities from pathology or radiology data[116], in chapter 3 we focus on capturing genotype-phenotype associations using neural networks as a means of connecting data views. By leveraging single cell data from a set of *IDH1/2* mutant AML patients, we develop deep learning approaches to explore how genotypic changes are reflected in cell specific gene expression signals. Specifically, this chapter examines if single cell gene expression patterns together with the computational power of neural networks have the capacity to predict a cell's status (malignant or WT) and subsequently its genotype in the context of *IDH1/2* mutated AML. Thus, first, we train a feedforward neural network to predict the cell's malignant or wild-type (WT) status in a binary fashion using single cell RNA sequencing data from 6 AML patients and 4 healthy individuals (50,026 cells in total). Then, in a multi-label setting, we train, for a single patient, a similar architecture to predict the mutational status of specific genomic abnormalities at the single cell level. In the hold-out test sets, the binary classification model achieved an accuracy of 98% while the multi-label one achieved a macro-average AUC ROC of 0.84. Additionally, the latter model showed notable efficiency (AUC ROC of 0.83) in predicting the subclonal *NRAS* mutational status, suggesting that *NRAS* mutations confer a distinctive gene expression pattern in *IDH1/2* AML. Concluding, chapter 3 applies deep learning to explore if and how single cell gene expression profiles can be predictive of the malignant cell status and the mutational profile of specific genetic abnormalities in *IDH1/2* AML in a supervised training context, and shows the potential of such modeling approaches in capturing meaningful genotype-phenotype relationships.

The work presented in Chapter 2 has been published here (Blood Advances journal) and the work presented in Chapter 3 has been published here (Association for Computing Machinery, International Conference on Bioscience, Biochemistry and Bioinformatics 2023 Conference proceedings).

1.7. References

1. Pan-cancer analysis of whole genomes. Nature. 2020;578: 82–93.

2. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2018;173: 1823.

3. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015;349: 1483–1489.

4. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458: 719–724.

5. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168: 613–628.

6. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481: 306–313.

7. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464: 993–998.

8. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45: 1113–1120.

9. Gerstung M, Papaemmanuil E, Martincorena I, Bullinger L, Gaidzik VI, Paschka P, et al. Precision oncology for acute myeloid leukemia using a knowledge bank approach. Nat Genet. 2017;49: 332–340.

10. Zhang N, Wu J, Wang Q, Liang Y, Li X, Chen G, et al. Global burden of hematologic malignancies and evolution patterns over the past 30 years. Blood Cancer J. 2023;13: 82.

11. Karagianni P, Giannouli S, Voulgarelis M. From the (Epi)Genome to Metabolism and Vice Versa; Examples from Hematologic Malignancy. Int J Mol Sci. 2021;22. doi:10.3390/ijms22126321

12. Maloy S, Hughes K. Brenner's Encyclopedia of Genetics. Academic Press; 2013.

13. Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. Cell. 2008;132: 631–644.

14. Crisan M, Dzierzak E. Correction: The many faces of hematopoietic stem cell heterogeneity. Development doi: 10.1242/dev.114231. Development. 2017;144: 4195.

15. Zhang Y, Gao S, Xia J, Liu F. Hematopoietic Hierarchy - An Updated Roadmap. Trends Cell Biol. 2018;28: 976–986.

16. Doulatov S, Notta F, Laurenti E, Dick JE. Hematopoiesis: a human perspective. Cell Stem Cell. 2012;10: 120–136.

17. Laurenti E, Göttgens B. From haematopoietic stem cells to complex differentiation landscapes. Nature. 2018;553: 418–426.

18. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48: 1193–1203.

19. Murati A, Brecqueville M, Devillier R, Mozziconacci M-J, Gelsi-Boyer V, Birnbaum D. Myeloid malignancies: mutations, models and management. BMC Cancer. 2012;12: 304.

20. Korn C, Méndez-Ferrer S. Myeloid malignancies and the microenvironment. Blood. 2017;129: 811–822.

21. Malcovati L, Ambaglio I, Elena C. The genomic landscape of myeloid neoplasms with myelodysplasia and its clinical implications. Curr Opin Oncol. 2015;27: 551–559.

22. Khoury JD, Solary E, Abla O, Akkari Y, Alaggio R, Apperley JF, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. Leukemia. 2022;36: 1703–1719.

23. Cazzola M. Myelodysplastic Syndromes. N Engl J Med. 2020;383: 1358–1374.

24. Marnell CS, Bick A, Natarajan P. Clonal hematopoiesis of indeterminate potential (CHIP): Linking somatic mutations, hematopoiesis, chronic inflammation and cardiovascular disease. J Mol Cell Cardiol. 2021;161: 98–105.

25. Bejar R. Implications of molecular genetic diversity in myelodysplastic syndromes. Curr Opin Hematol. 2017;24: 73–78.

26. Current standard of care in patients with myelodysplastic syndromes and future perspectives. healthbook TIMES Onco Hema. 2020. doi:10.36000/hbt.oh.2020.06.026

27. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood. 2013;122: 3616–27; quiz 3699.

28. Haferlach T, Nagata Y, Grossmann V, Okuno Y, Bacher U, Nagae G, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia. 2014;28: 241–247.

29. Visconte V, Tiu RV, Rogers HJ. Pathogenesis of myelodysplastic syndromes: an overview of molecular and non-molecular aspects of the disease. Blood Res. 2014;49: 216–227.

30. Bernard E, Nannya Y, Hasserjian RP, Devlin SM, Tuechler H, Medina-Martinez JS, et al. Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes. Nat Med. 2020;26: 1549–1556.

31. Bernard E, Tuechler H, Greenberg PL, Hasserjian RP, Arango Ossa JE, Nannya Y, et al. Molecular international prognostic scoring system for myelodysplastic syndromes. NEJM Evid. 2022;1. doi:10.1056/evidoa2200008

32. Malcovati L, Stevenson K, Papaemmanuil E, Neuberg D, Bejar R, Boultwood J, et al. SF3B1-mutant MDS as a distinct disease subtype: a proposal from the International Working Group for the Prognosis of MDS. Blood. 2020;136: 157–170.

33. Greenberg PL, Tuechler H, Schanz J, Sanz G, Garcia-Manero G, Solé F, et al. Revised international prognostic scoring system for myelodysplastic syndromes. Blood. 2012;120: 2454–2465.

34. Kontandreopoulou C-N, Kalopisis K, Viniou N-A, Diamantopoulos P. The genetics of myelodysplastic syndromes and the opportunities for tailored treatments. Front Oncol. 2022;12: 989483.

35. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med. 2011;365: 1384–1395.

36. Kataoka N, Matsumoto E, Masaki S. Mechanistic Insights of Aberrant Splicing with Splicing Factor Mutations Found in Myelodysplastic Syndromes. Int J Mol Sci. 2021;22. doi:10.3390/ijms22157789

37. Taylor J, Lee SC. Mutations in spliceosome genes and therapeutic opportunities in myeloid malignancies. Genes Chromosomes Cancer. 2019;58: 889–902.

38. Pellagatti A, Armstrong RN, Steeples V, Sharma E, Repapi E, Singh S, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood. 2018;132: 1225–1240.

39. Menssen AJ, Walter MJ. Genetics of progression from MDS to secondary leukemia. Blood. 2020;136: 50–60.

40. Capelli D, Menotti D, Fiorentini A, Saraceni F, Olivieri A. Secondary Acute Myeloid Leukemia: Pathogenesis and Treatment. In: Li W, editor. Leukemia. Brisbane (AU): Exon Publications;

41. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. N Engl J Med. 2016;374: 2209–2221.

42. Döhner H, Weisdorf DJ, Bloomfield CD. Acute Myeloid Leukemia. N Engl J Med. 2015;373: 1136–1152.

43. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017;129: 424–447.

44. DiNardo CD, Erba HP, Freeman SD, Wei AH. Acute myeloid leukaemia. Lancet. 2023;401: 2073–2086.

45. Abelson S, Collord G, Ng SWK, Weissbrod O, Mendelson Cohen N, Niemeyer E, et al. Prediction of acute myeloid leukaemia risk in healthy individuals. Nature. 2018;559: 400–404.

46. Stubbins RJ, Francis A, Kuchenbauer F, Sanford D. Management of Acute Myeloid Leukemia: A Review for General Practitioners in Oncology. Curr Oncol. 2022;29: 6245–6259.

47. De Kouchkovsky I, Abdul-Hay M. Acute myeloid leukemia: a comprehensive review and 2016 update. Blood Cancer J. 2016;6: e441.

48. Lindsley RC, Mar BG, Mazzola E, Grauman PV, Shareef S, Allen SL, et al. Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. Blood. 2015;125: 1367–1376.

49. Kantarjian H, Kadia T, DiNardo C, Daver N, Borthakur G, Jabbour E, et al. Acute myeloid leukemia: current progress and future directions. Blood Cancer J. 2021;11: 41.

50. Figueroa ME, Abdel-Wahab O, Lu C, Ward PS, Patel J, Shih A, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell. 2010;18: 553–567.

51. Xu W, Yang H, Liu Y, Yang Y, Wang P, Kim S-H, et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. Cancer Cell. 2011;19: 17–30.

52. Lu C, Ward PS, Kapoor GS, Rohle D, Turcan S, Abdel-Wahab O, et al. IDH mutation impairs histone demethylation and results in a block to cell differentiation. Nature. 2012;483: 474–478.

53. Kats LM, Reschke M, Taulli R, Pozdnyakova O, Burgess K, Bhargava P, et al. Proto-oncogenic role of mutant IDH2 in leukemia initiation and maintenance. Cell Stem Cell. 2014;14: 329–341.

54. Prada-Arismendy J, Arroyave JC, Röthlisberger S. Molecular biomarkers in acute myeloid leukemia. Blood Rev. 2017;31: 63–76.

55. Stein EM, DiNardo CD, Pollyea DA, Fathi AT, Roboz GJ, Altman JK, et al. Enasidenib in mutant relapsed or refractory acute myeloid leukemia. Blood. 2017;130: 722–731.

56. DiNardo CD, Stein EM, de Botton S, Roboz GJ, Altman JK, Mims AS, et al. Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. N Engl J Med. 2018;378: 2386–2398.

57. Stein EM, DiNardo CD, Fathi AT, Mims AS, Pratz KW, Savona MR, et al. Ivosidenib or enasidenib combined with intensive chemotherapy in patients with newly diagnosed AML: a phase 1 study. Blood. 2021;137: 1792–1803.

58. Choe S, Wang H, DiNardo CD, Stein EM, de Botton S, Roboz GJ, et al. Molecular mechanisms mediating relapse following ivosidenib monotherapy in IDH1-mutant relapsed or refractory AML. Blood Adv. 2020;4: 1894–1905.

59. Visconte V, Tabarroki A, Zhang L, Parker Y, Hasrouni E, Mahfouz R, et al. Splicing factor 3b subunit 1 (Sf3b1) haploinsufficient mice display features of low risk Myelodysplastic syndromes with ring sideroblasts. J Hematol Oncol. 2014;7: 89.

60. Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, et al. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. Cancer Cell. 2016;30: 404–417.

61. Skopek R, Palusińska M, Kaczor-Keller K, Pingwara R, Papierniak-Wyglądała A, Schenk T, et al. Choosing the Right Cell Line for Acute Myeloid Leukemia (AML) Research. Int J Mol Sci. 2023;24. doi:10.3390/ijms24065377

62. Kurkowiak M, Pępek M, Machnicki MM, Solarska I, Borg K, Rydzanicz M, et al. Genomic landscape of human erythroleukemia K562 cell line, as determined by next-generation sequencing and cytogenetics. Acta Haematol Pol. 2017;48: 343–349.

63. Kotini AG, Chang C-J, Chow A, Yuan H, Ho T-C, Wang T, et al. Stage-Specific Human Induced Pluripotent Stem Cells Map the Progression of Myeloid Transformation to Transplantable Leukemia. Cell Stem Cell. 2017;20: 315–328.e7.

64. Kotini AG, Carcamo S, Cruz-Rodriguez N, Olszewska M, Wang T, Demircioglu D, et al. Patient-Derived iPSCs Faithfully Represent the Genetic Diversity and Cellular Architecture of Human Acute Myeloid Leukemia. Blood cancer discovery. 2023. pp. 318–335.

65. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007;131: 861–872.

66. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126: 663–676.

67. Bodein A, Scott-Boyer M-P, Perin O, Lê Cao K-A, Droit A. Interpretation of network-based integration from multi-omics longitudinal data. Nucleic Acids Res. 2022;50: e27.

68. Sun YV, Hu Y-J. Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. Adv Genet. 2016;93: 147–190.

69. He X, Liu X, Zuo F, Shi H, Jing J. Artificial intelligence-based multi-omics analysis fuels cancer precision medicine. Semin Cancer Biol. 2023;88: 187–200.

70. Wang Q, Peng W-X, Wang L, Ye L. Toward multiomics-based next-generation diagnostics for precision medicine. Per Med. 2019;16: 157–170.

71. Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J. 2021;19: 949–960.

72. Mohorianu I, Bretman A, Smith DT, Fowler EK, Dalmay T, Chapman T. Comparison of alternative approaches for analysing multi-level RNA-seq data. PLoS One. 2017;12: e0182694.

73. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat Protoc. 2022;17: 1518–1552.

74. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience. 2022;25: 103798.

75. Wu J, Li Y, Feng D, Yu Y, Long H, Hu Z, et al. Integrated analysis of ATAC-seq and RNA-seq reveals the transcriptional regulation network in SLE. Int Immunopharmacol. 2023;116: 109803.

76. Fan J, Lee H-O, Lee S, Ryu D-E, Lee S, Xue C, et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res. 2018;28: 1217–1227.

77. Wang X, Wu X, Hong N, Jin W. Progress in single-cell multimodal sequencing and multi-omics data integration. Biophys Rev. 2023. doi:10.1007/s12551-023-01092-3

78. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. Exp Mol Med. 2020;52: 1428–1442.

79. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9: 5233.

80. Žurauskienė J, Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics. 2016;17: 140.

81. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14: 483–486.

82. van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018;174: 716–729.e27.

83. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun. 2018;9: 997.

84. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics. 2017;33: 2314–2321.

85. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14: 1083–1086.

86. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13: 845–848.

87. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019;20: 59.

88. Kashima Y, Sakamoto Y, Kaneko K, Seki M, Suzuki Y, Suzuki A. Single-cell sequencing techniques from individual to multiomics analyses. Exp Mol Med. 2020;52: 1419–1427.

89. Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat Methods. 2017;14: 491–493.

90. Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. Monovar: single-nucleotide variant detection in single cells. Nat Methods. 2016;13: 505–507.

91. Bohrson CL, Barton AR, Lodato MA, Rodin RE, Luquette LJ, Viswanadham VV, et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat Genet. 2019;51: 749–754.

92. Momeni Z, Hassanzadeh E, Saniee Abadeh M, Bellazzi R. A survey on single and multi omics data mining methods in cancer data classification. J Biomed Inform. 2020;107: 103466.

93. Heo YJ, Hwa C, Lee G-H, Park J-M, An J-Y. Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes. Mol Cells. 2021;44: 433–443.

94. Mitra S, Saha S, Hasanuzzaman M. Multi-view clustering for multi-omics data using unified embedding. Sci Rep. 2020;10: 13654.

95. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16: 85–97.

96. Agamah FE, Bayjanov JR, Niehues A, Njoku KF, Skelton M, Mazandu GK, et al. Computational approaches for network-based integrative multi-omics analysis. Front Mol Biosci. 2022;9: 967205.

97. Ha KCH, Sterne-Weiler T, Morris Q, Weatheritt RJ, Blencowe BJ. Differential contribution of transcriptomic regulatory layers in the definition of neuronal identity. Nat Commun. 2021;12: 335.

98. Hua X, Wang Y-Y, Jia P, Xiong Q, Hu Y, Chang Y, et al. Multi-level transcriptome sequencing identifies COL1A1 as a candidate marker in human heart failure progression. BMC Med. 2020;18: 2.

99. Arjmand B, Hamidpour SK, Tayanloo-Beik A, Goodarzi P, Aghayan HR, Adibi H, et al. Machine Learning: A New Prospect in Multi-Omics Data Analysis of Cancer. Front Genet. 2022;13: 824451.

100. Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. Front Oncol. 2020;10: 1030.

101. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022;23. doi:10.1093/bib/bbab454

102. Xiang G, Keller CA, Heuston E, Giardine BM, An L, Wixom AQ, et al. An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis. Genome Res. 2020;30: 472–484.

103. Chen X, Wang P, Qiu H, Zhu Y, Zhang X, Zhang Y, et al. Integrative epigenomic and transcriptomic analysis reveals the requirement of JUNB for hematopoietic fate induction. Nat Commun. 2022;13: 3131.

104. Adossa N, Khan S, Rytkönen KT, Elo LL. Computational strategies for single-cell multi-omics integration. Comput Struct Biotechnol J. 2021;19: 2588–2596.

105. Han KY, Kim K-T, Joung J-G, Son D-S, Kim YJ, Jo A, et al. SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. Genome Res. 2018;28: 75–87.

106. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015;12: 519–522.

107. Rodriguez-Meira A, O'Sullivan J, Rahman H, Mead AJ. TARGET-Seq: A Protocol for High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. STAR Protoc. 2020;1: 100125.

108. Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. Nat Commun. 2020;11: 89.

109. Muyas F, Sauer CM, Valle-Inclán JE, Li R, Rahbari R, Mitchell TJ, et al. De novo detection of somatic mutations in high-throughput single-cell profiling data sets. Nat Biotechnol. 2023. doi:10.1038/s41587-023-01863-z

110. Dou J, Tan Y, Kock KH, Wang J, Cheng X, Tan LM, et al. Single-nucleotide variant calling in single-cell sequencing data with Monopogen. Nat Biotechnol. 2023. doi:10.1038/s41587-023-01873-x

111. Website. Available: inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV

112. Nam AS, Kim K-T, Chaligne R, Izzo F, Ang C, Taylor J, et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. Nature. 2019;571: 355–360.

113. Petti AA, Williams SR, Miller CA, Fiddes IT, Srivatsan SN, Chen DY, et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. Nat Commun. 2019;10: 3660.

114. Rodriguez-Meira A, Buck G, Clark S-A, Povinelli BJ, Alcolea V, Louka E, et al. Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. Mol Cell. 2019;73: 1292–1305.e8.

115. van Galen P, Hovestadt V, Wadsworth MH Ii, Hughes TK, Griffin GK, Battaglia S, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell. 2019;176: 1265–1281.e24.

116. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. Cancer Cell. 2022;40: 1095–1110.

117. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005;37: 710–717.

118. Hartford CM, Duan S, Delaney SM, Mi S, Kistner EO, Lamba JK, et al. Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. Blood. 2009;113: 2145–2153.

119. Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, Clark TA, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci U S A. 2007;104: 9758–9763.

120. Huang RS, Duan S, Kistner EO, Hartford CM, Dolan ME. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. Mol Cancer Ther. 2008;7: 3038–3046.

121. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306: 636–640.

122. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30.

123. Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501: 506–511.

124. Maynard ND, Chen J, Stuart RK, Fan J-B, Ren B. Genome-wide mapping of allele-specific protein-DNA interactions in human cells. Nat Methods. 2008;5: 307–309.

125. Fridley BL, Lund S, Jenkins GD, Wang L. A Bayesian integrative genomic model for pathway analysis of complex traits. Genet Epidemiol. 2012;36: 352–359.

126. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS One. 2011;6: e24709.

127. Tao M, Song T, Du W, Han S, Zuo C, Li Y, et al. Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data. Genes . 2019;10. doi:10.3390/genes10030200

128. Aure MR, Vitelli V, Jernström S, Kumar S, Krohn M, Due EU, et al. Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. Breast Cancer Res. 2017;19: 44.

129. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490: 61–70.

130. Huh R, Yang Y, Jiang Y, Shen Y, Li Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. Nucleic Acids Res. 2020;48: 86–95.

131. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11: 333–337.

132. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12: 3445.

133. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun. 2018;9: 4453.

134. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184: 3573–3587.e29.

135. Singh R, Hie BL, Narayan A, Berger B. Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. Genome Biol. 2021;22: 131.

136. Kim HJ, Lin Y, Geddes TA, Yang JYH, Yang P. CiteFuse enables multi-modal analysis of CITE-seq data. Bioinformatics. 2020;36: 4137–4143.

137. Lock EF, Hoadley KA, Marron JS, Nobel AB. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. Ann Appl Stat. 2013;7: 523–542.

138. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14: e8124.

139. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell. 2019;177: 1873–1887.e17.

140. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2016;32: 1–8.

141. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol. 2009;8: Article28.

142. Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. Mach Learn. 2011;83: 331–353.

143. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177: 1888–1902.e21.

144. Wang C, Sun D, Huang X, Wan C, Li Z, Han Y, et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol. 2020;21: 198.

145. Yu T. AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. PLoS Comput Biol. 2022;18: e1009826.

146. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods. 2021;18: 272–282.

147. Lin X, Tian T, Wei Z, Hakonarson H. Clustering of single-cell multi-omics data with a multimodal deep learning method. Nat Commun. 2022;13: 7705.

148. Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013;29: 2610–2616.

149. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28: 3290–3297.

150. Kormaksson M, Booth JG, Figueroa ME, Melnick A. Integrative model-based clustering of microarray methylation and expression data. Ann. Appl. Stat. 2012;6 (3) 1327 - 1347

151. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25: 2906–2912.

152. Wang X, Sun Z, Zhang Y, Xu Z, Xin H, Huang H, et al. BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data. Nucleic Acids Res. 2020;48: 5814–5824.

153. Campbell KR, Steif A, Laks E, Zahn H, Lai D, McPherson A, et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. 2019;20: 54.

154. Sanghi A, Gruber JJ, Metwally A, Jiang L, Reynolds W, Sunwoo J, et al. Chromatin accessibility associates with protein-RNA correlation in human cancer. Nat Commun. 2021;12: 5732.

155. Wang T, Pine AR, Kotini AG, Yuan H, Zamparo L, Starczynowski DT, et al. Sequential CRISPR gene editing in human iPSCs charts the clonal evolution of myeloid leukemia and identifies early disease targets. Cell Stem Cell. 2021;28: 1074–1089.e7.

156. Wesely J, Kotini AG, Izzo F, Luo H, Yuan H, Sun J, et al. Acute Myeloid Leukemia iPSCs Reveal a Role for RUNX1 in the Maintenance of Human Leukemia Stem Cells. Cell Rep. 2020;31: 107688.

157. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020;11: 3877.

158. Fotis C, Meimetis N, Sardis A, Alexopoulos LG. DeepSIBA: chemical structure-based inference of biological alterations using deep learning. Mol Omics. 2021;17: 108–120.

159. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24: 1559–1567.

160. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. doi:10.1109/cvpr.2016.308

161. Brück OE, Lallukka-Brück SE, Hohtari HR, Ianevski A, Ebeling FT, Kovanen PE, et al. Machine Learning of Bone Marrow Histopathology Identifies Genetic and Clinical Determinants in Patients with MDS. Blood cancer discovery. 2021. pp. 238–249.

162. Chen M, Zhang B, Topatana W, Cao J, Zhu H, Juengpanich S, et al. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. NPJ Precis Oncol. 2020;4: 14.

163. Loeffler CML, Ortiz Bruechle N, Jung M, Seillier L, Rose M, Laleh NG, et al. Artificial Intelligence-based Detection of FGFR3 Mutational Status Directly from Routine Histology in Bladder Cancer: A Possible Preselection for Molecular Testing? Eur Urol Focus. 2022;8: 472–479.

164. Jang H-J, Lee A, Kang J, Song IH, Lee SH. Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. World J Gastroenterol. 2020;26: 6207–6223.

165. Tsou P, Wu C-J. Mapping Driver Mutations to Histopathological Subtypes in Papillary Thyroid Carcinoma: Applying a Deep Convolutional Neural Network. J Clin Med Res. 2019;8. doi:10.3390/jcm8101675

166. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer. 2020;1: 800–810.

167. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer. 2020;1: 789–799.

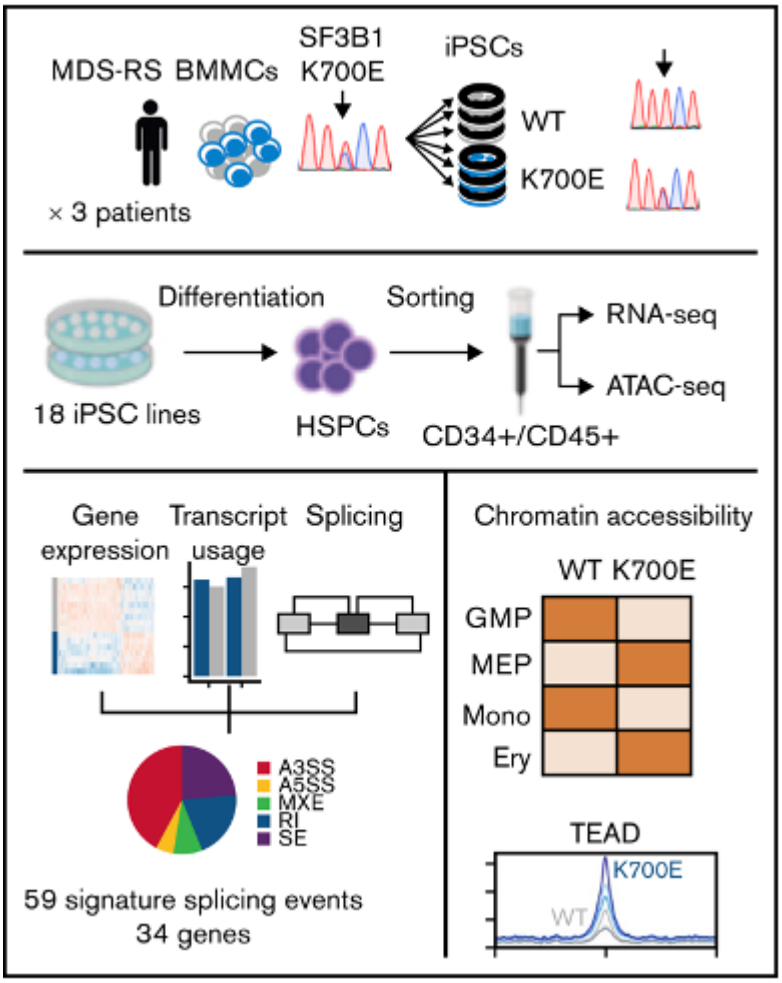
168. Ha SM, Chae EY, Cha JH, Kim HH, Shin HJ, Choi WJ. Association of BRCA Mutation Types, Imaging Features, and Pathologic Findings in Patients With Breast Cancer With BRCA1 and BRCA2 Mutations. AJR Am J Roentgenol. 2017;209: 920–928.

169. Wang S, Shi J, Ye Z, Dong D, Yu D, Zhou M, et al. Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur Respir J. 2019;53. doi:10.1183/13993003.00986-2018

170. He K, Liu X, Li M, Li X, Yang H, Zhang H. Noninvasive KRAS mutation estimation in colorectal cancer using a deep learning method based on CT imaging. BMC Med Imaging. 2020;20: 59.

Chapter 2

Patient-specific MDS-RS iPSCs define the mis-spliced transcript repertoire and chromatin landscape of *SF3B1*-mutant HSPCs



2.1. Chapter abstract

SF3B1^{K700E} is the most frequent mutation in myelodysplastic syndrome (MDS), but the mechanisms by which it drives MDS pathogenesis remain unclear. We harnessed a panel of 18 genetically matched *SF3B1*^{K700E}- and *SF3B1*^{WT}-induced pluripotent stem cell (iPSC) lines from patients with MDS with ring sideroblasts (MDS-RS) harboring isolated *SF3B1*^{K700E} mutations. RNA and ATAC sequencing was performed in purified CD34^{+}/CD45^{+} hematopoietic stem/progenitor cells (HSPCs) derived from these lines. We developed a novel computational framework integrating splicing with transcript usage and gene expression analyses and derived a *SF3B1*^{K700E} splicing signature consisting of 59 splicing events linked to 34 genes, which associates with the *SF3B1* mutational status of primary MDS patient cells. The chromatin landscape of *SF3B1*^{K700E} HSPCs showed increased priming toward the megakaryocyte-erythroid lineage. Transcription factor (TF) motifs enriched in chromatin regions more accessible in *SF3B1*^{K700E} cells included, unexpectedly, motifs of the TEA domain (TEAD) transcription factor family. TEAD expression and transcriptional activity were upregulated in *SF3B1*-mutant iPSC-HSPCs, in support of a Hippo pathway-independent role of TEAD as a potential novel transcriptional regulator of *SF3B1*^{K700E} cells. This study provides a comprehensive characterization of the transcriptional and chromatin landscape of *SF3B1*^{K700E} HSPCs and nominates mis-spliced genes and transcriptional programs with putative roles in MDS-RS disease biology

2.2. Introduction

Myelodysplastic syndromes (MDS) are myeloid malignancies characterized by ineffective hematopoiesis, blood cytopenias, and an increased risk of progression to secondary acute myeloid leukemia[1]. Recurrent somatic mutations in genes encoding splicing factors (SFs) were discovered a decade ago as a novel class of driver mutations in MDS, collectively occurring in more than 50% of patients with MDS[2–5]. Mutations in splicing factor 3B, subunit 1 (*SF3B1*), are present in ~ 24% of patients with MDS and define a distinct MDS clinical subgroup, termed MDS with ring sideroblasts (MDS-RS), characterized by erythroblasts with abnormal iron accumulation in mitochondria that form a ring around the cell nucleus (ring sideroblasts), ineffective erythropoiesis, macrocytic anemia, and favorable prognosis[3–7].

SF3B1 is a core spliceosomal protein (a key component of the U2 small nuclear ribonucleoprotein complex [snRNP]) that binds upstream of the branch point and is required to facilitate 3’ splice site recognition of most introns[8]. Nearly all mutations in *SF3B1* are heterozygous, most commonly target the K700 hotspot, and result in altered RNA-binding specificity of mutant SF3B1. *SF3B1* mutations are associated with preferential use of cryptic 3’ splice sites, leading to nonsense-mediated decay (NMD) or generation of different isoforms of multiple transcripts[9–11].

Some recent studies implicated specific mis-splicing events associated with *SF3B1* mutations in the pathogenesis of MDS or other malignancies. An alternative erythroferrone (*ERFE*) transcript in *SF3B1*-mutant erythroid lineage cells was linked to disruption of iron homeostasis[12]. Decrease in expression of *BRD9*, a component of the noncanonical BRG1-associated factors (BAF) chromatin-remodeling complex, through

inclusion of a “poison exon”, was also shown to confer oncogenic properties in uveal melanoma models[13]. Despite these insights, the mechanisms by which mutant SF3B1 drives MDS, and malignancy in general, remain incompletely understood, and the critical mis-splicing events that mediate these effects are not well characterized. Importantly, mis-splicing events have thus far been cataloged either in primary patient cells or murine or cellular models, each with distinct limitations. Patient samples are heterogeneous in terms of clonality, presence of co-occurring mutations, and cell type composition. Conversely, murine models have the important limitation that alternative splicing events are largely non conserved between mouse and human[14]. Finally, previous cellular models of *SF3B1* mutations consisted of engineered immortalized leukemia cell lines (such as K562), which harbor mutations not related to MDS pathogenesis, and result in abnormal levels and stoichiometry of mutant and wild-type (WT) SF3B1 because of aneuploidy and/or use of overexpression systems.

Here, we leveraged a panel of karyotypically normal diploid-induced pluripotent stem cell (iPSC) lines with an isolated *SF3B1*^{K700E} mutation, as well as genetically matched WT iPSCs, from patients with MDS-RS. By integrating splicing, gene expression, and transcript usage analyses, we derived a splicing signature of mutant SF3B1 that we validated in datasets of patients with MDS. Furthermore, we characterized the chromatin landscape of *SF3B1*^{K700E} iPSC-derived hematopoietic stem/progenitor cells (iPSC-HSPCs) and identified increased transcriptional activity of the TEAD family of transcription factors (TFs) in mutant cells. This study provides a refined view of the altered misspliced transcriptome of human *SF3B1*^{K700E} HSPCs and characterizes for the first time their chromatin landscape, pinpointing TEAD as a potential regulator of *SF3B1*^{K700E} HSPCs.

2.3. Data & Methods

From a previous population genome profiling study[3], 3 BM mononuclear cell samples from 3 patients with MDS-RS (P21, P22, P23) were identified. These patients harbored isolated *SF3B1*^{K700E} mutations with high variant allele frequencies (VAFs; range, 37%-42%; Figure 2.1; supplemental Table 2.1). Upon cell reprogramming, both *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC lines from all patients were obtained. Specifically, 3 independent *SF3B1*^{K700E} and 3 *SF3B1*^{WT} iPSC lines from each patient (total lines = 18) were established, serving as biological replicates (supplemental Table 2.2). The presence of any karyotypic abnormalities in any of these lines was excluded. The presence of any other MDS/AML driver mutations in all lines or in the starting cells was also excluded by using next generation sequencing of a panel of 126 genes implicated in myeloid malignancy[15]. For these iPSC lines, directed hematopoietic differentiation was performed using protocols[16] from the Papapetrou laboratory and CD34^{+}/CD45^{+} HSPCs were collected for RNA and ATAC-sequencing. (We note that the data generation process including the derivation and differentiation of the iPSC lines is not part of the current thesis).

RNA-sequencing analysis

RNA was extracted with the Direct-zol RNA purification kit (Zymo R2061). Sequencing libraries were prepared using the TruSeq Stranded mRNA library prep kit (Illumina 20020594) from 500 ng input RNA. Samples were

barcoded and run on a Hi-seq 4000 in a 100-bp/100-bp paired-end run, using the Hi-seq 3000/4000 SBS kit (Illumina).

HSPC samples from 16 iPSC lines were included in the RNA-seq analyses after quality control of the raw data (supplemental Table 2.2). RNA-seq reads from the fastq files were mapped to the GRCh37 assembly of the human genome using the STAR aligner[17]. The Ensembl GRCh37 gene and transcript annotations were used. Salmon[18] was used to perform transcript quantification, and gene counts were generated from the transcript level abundances using the tximport function of the tximport R package[19]. Differential gene expression analysis was performed using DESeq2[20]. Genes with a false discovery rate (FDR) < 0.05 and absolute expression $\log_2$fc > 1 in *SF3B1*^{K700E} vs *SF3B1*^{WT} cells were considered as differentially expressed.

Differential transcript usage between the *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs was performed using the DEXSeq[21] and stageR[22] R-packages. Transcripts with a relative abundance proportion <5% in all samples were filtered. Transcripts were considered to have differential usage if absolute usage $\log_2$fc was >1 and overall FDR was < 0.05 (supplemental Methods).

Differential alternative splicing was performed using the rMATS[23] tool using the aligned BAM files. The relative expression (inclusion level) of alternatively spliced isoforms was estimated by the fraction of reads mapping to an alternative splicing event over the total reads[23]. Events with FDR < 0.05 and absolute inclusion level difference >10% were considered as differentially spliced between the *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs.

Integration framework of differential gene expression, transcript usage, and splicing

To generate an *SF3B1*^{K700E} signature, we combined differential gene expression, differential transcript usage, and differential splicing analyses in a multi-stage fusion setting[24,25]. First, we identified the set of transcripts that contain the exons present in each differential splicing event using the maser R-package[26]. We then filtered out non-differentially used transcripts and paired each differential splicing event with the remaining set of differentially used transcripts. The pairs that belonged to genes with a statistically significant expression $\log_2$fc and contained a differential splicing event with an FDR value within the 20 lowermost FDR values were considered as the “tier 1” set, from which the mutant SF3B1 signature events and genes were derived (supplemental Methods).

ATAC-sequencing analysis

Nuclear pellets (supplemental Methods) were subjected to transposase reaction using the Illumina Nextera DNA Sample Preparation Kit. The libraries were quantified using the Agilent BioAnalyzer. Sequencing of 75 nucleotide-long paired-end reads was performed in a NextSeq-500 (Illumina).

HSPC samples from 15 iPSC lines were included in the ATAC-seq analyses after quality control of the raw data (supplemental Table 2.2). ATAC-seq reads from the fastq files were trimmed with the TrimGalore tool to

remove adaptor sequences and then aligned to the GRCh37 reference genome using the Bowtie2[27] aligner. Reads with a mapping quality (MAPQ) score < 10 were removed using samtools. Duplicate reads were removed using Picard. All aligned reads were shifted to remove Tn5 transposase artifacts, as previously described[28] using deeptools[29]. Peaks were called using MACS2[30] (supplemental Methods) and then filtered using the irreproducible discovery rate[30,31] framework with a cutoff of 0.05. Then, we merged all reproducible peaks to create an ATAC-seq atlas. Differential accessibility analysis was performed using DESeq2. Peaks with an FDR cutoff of 0.05 and absolute $\log_2$fc > 1 were considered differentially accessible.

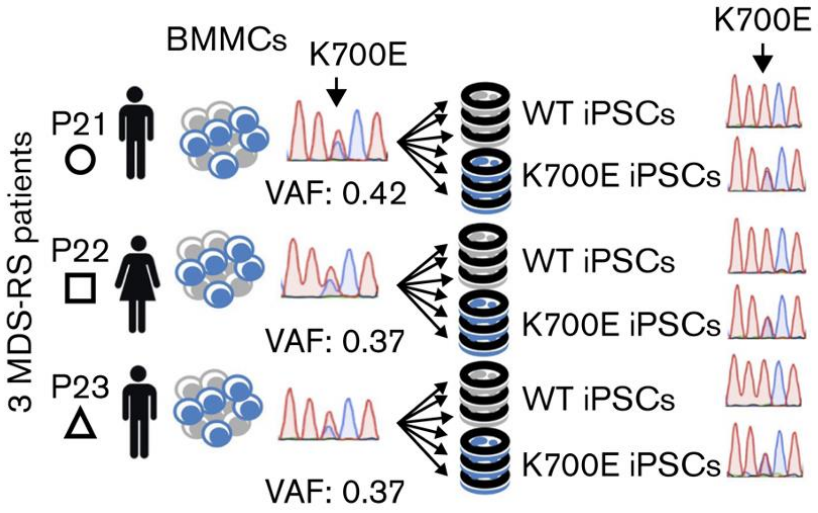


Figure 2.1. Schematic overview of the derivation of iPSC lines with isolated *SF3B1*^{K700E} mutation and genetically matched normal WT lines from 3 patients with MDS-RS (BMMCs, bone marrow mononuclear cells).

2.4. Results

Global gene expression, mis-splicing, and differential transcript usage in *SF3B1*^{K700E} HSPCs

To examine the effects of *SF3B1*^{K700E} in the transcriptome, we performed RNA sequencing in sorted CD34$^+$/CD45$^+$ iPSC-HSPCs from 3 *SF3B1*^{K700E} and 3 *SF3B1*^{WT} iPSC lines from each patient (total 18 lines; supplemental Table 2.2). Samples MDS-22.1 and MDS-22.43 did not pass quality control at the library preparation stage and were not included in the analyses. Principal component analysis (PCA) and hierarchical clustering based on gene expression grouped the iPSC lines by genotype (i.e. *SF3B1*^{K700E} vs *SF3B1*^{WT}; Figure 2.2A-B).

Differential gene expression analysis revealed 2737 differentially expressed genes in the *SF3B1*^{K700E} mutant vs WT lines, 1821 of which were upregulated in the *SF3B1*^{K700E} cells (supplemental Figure 2.1A-B). Gene set enrichment analysis showed enrichment of gene sets related to metabolism and cell morphology in genes

upregulated in *SF3B1*^{K700E} cells and enrichment of genes related to myeloid lineage differentiation in the downregulated genes (supplemental Figure 2.1C-E).

To examine the effects of the *SF3B1*^{K700E} mutation on splicing, we characterized alternative splicing (AS) events in the *SF3B1*^{K700E} and *SF3B1*^{WT} cells, classified as alternative 3' splice site use (A3SS), alternative 5' splice site use (A5SS), mutually exclusive exons (MXE), retention of introns (RI), and skipping (inclusion or exclusion) of cassette exons (SE). A total of 1829 differential splicing events were detected between *SF3B1*^{K700E} and *SF3B1*^{WT} cells, which included 983 SE, 338 MXE, 265 A3SS, 173 RI, and 70 A5SS events (supplemental Figure 2.1F). Hierarchical clustering, as well as PCA, based on the inclusion levels of the differential splicing events, also separated the cells based on genotype, as expected (Figure 2.2C; supplemental Figure 2.1G). Consistent with previous studies, we found increased exclusion of cassette exons, increased use of alternative 3' splice sites, and decreased retention of introns in *SF3B1*^{K700E} cells (Figure 2.2D; supplemental Figure 2.1H)[9,11].

To evaluate the effects of the *SF3B1*^{K700E} mutation at the transcript level, we performed differential transcript usage analysis, which identified 1086 differentially used transcripts between *SF3B1*^{K700E} and *SF3B1*^{WT} cells (547 more used and 539 less used in *SF3B1*^{K700E} compared with *SF3B1*^{WT} cells). These differentially used transcripts belong to 865 genes, 198 of which were also found to be differentially expressed (supplemental Figure 2.1I). In summary, these analyses demonstrate that *SF3B1*^{K700E} mutations are associated with distinct gene expression, splicing, and transcript usage signatures.

Integration framework categorizes mutant SF3B1 gene targets by linking differential splicing to differential transcript usage and differential gene expression

Most previous studies have prioritized candidate target genes of mis-splicing by mutant SF3B1 in cancer cells, by selecting splicing events based on the size of differences in inclusion level of the isoforms between mutant and control cells[11]. To categorize splicing effects of the *SF3B1*^{K700E} mutation in MDS, we developed a computational multi-stage approach combining analyses at 3 different transcriptomic levels: gene expression, splicing, and transcript usage. This framework was used to classify the splicing events into 5 tier-based classes (supplemental Methods, supplemental Figure 2.2A; supplemental Table 2.3).

Of 1829 total differential splicing events between *SF3B1*^{K700E} and *SF3B1*^{WT} HSPCs, 215 were associated with at least 1 differentially used transcript. Of these 215 events, 95 belong to genes with a statistically significant (FDR < 0.05) expression $\log_2$ fold change ($\log_2$fc) between *SF3B1*^{K700E} and *SF3B1*^{WT} HSPCs. Of these 95 events, we selected the top 59 differentially spliced events (with the lowest 20 FDR values). These tier 1 59 events belong to 34 genes: 19 downregulated and 15 upregulated in *SF3B1*^{K700E} vs *SF3B1*^{WT} cells (Figures 2.2E and 2.3; supplemental Figure 2.2B; supplemental Table 2.4). Fifty-one (86%) of these 59 tier 1 events are A3SS, RI, or SE events (supplemental Figure 2.2C). This set of 59 events contained more A3SS events with increased use in *SF3B1*^{K700E} vs *SF3B1*^{WT} cells and more RI events that were less retained in *SF3B1*^{K700E} vs *SF3B1*^{WT} cells, reflecting the event distribution among all differential splicing events (Figure 2.3).

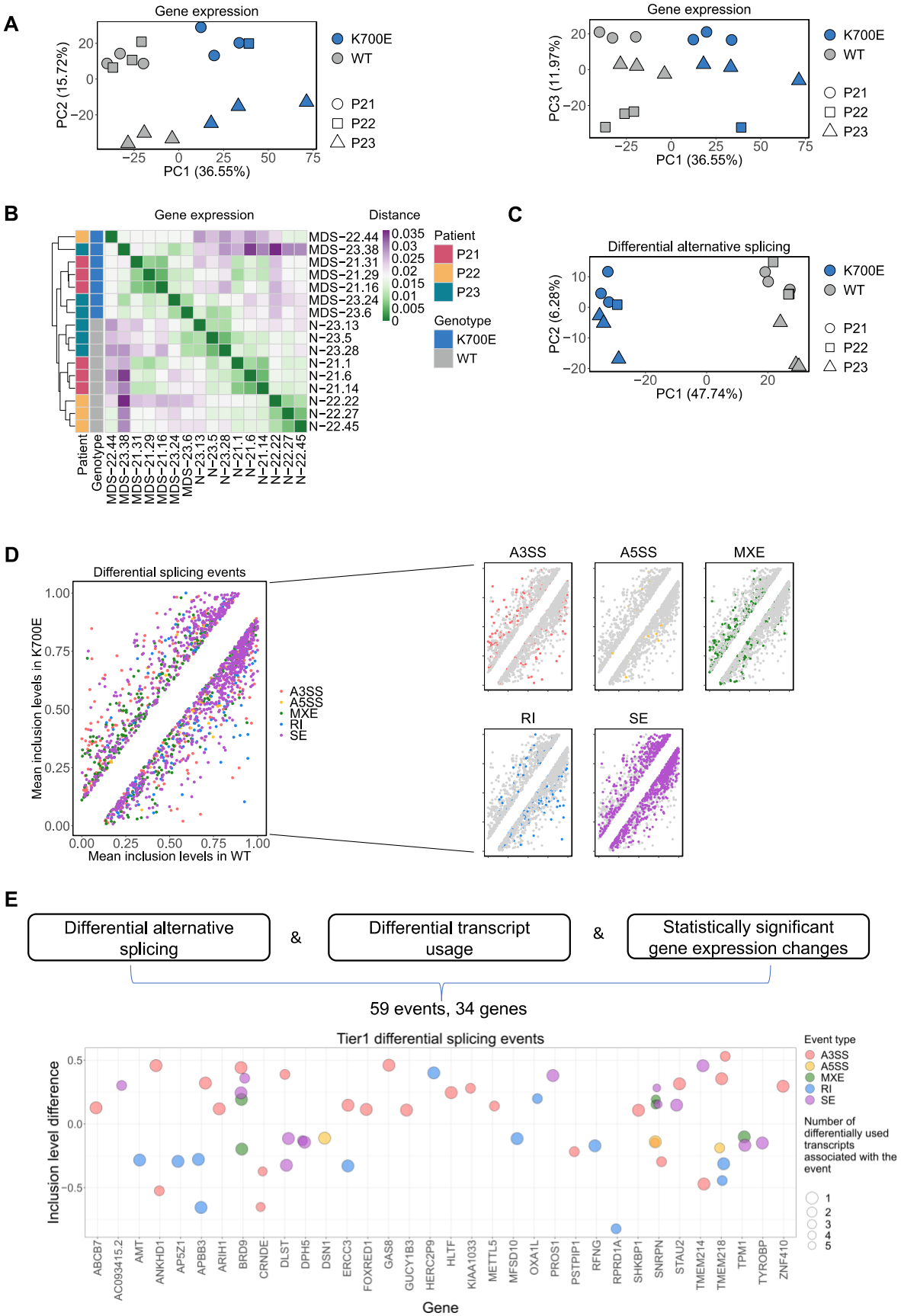

A
Gene expression
PC2 (15.72%)
PC1 (36.55%)
K700E
WT
P21
P22
P23
Gene expression
PC3 (11.97%)
PC1 (36.55%)
B
Gene expression
Distance
Patient
Genotype
MDS−22.44
MDS−23.38
MDS−21.31
MDS−21.29
MDS−21.16
MDS−23.24
MDS−23.6
N−23.13
N−23.5
N−23.28
N−21.1
N−21.6
N−21.14
N−22.22
N−22.27
N−22.45
C
Differential alternative splicing
PC2 (6.28%)
PC1 (47.74%)
D
Differential splicing events
Mean inclusion levels in K700E
Mean inclusion levels in WT
A3SS
A5SS
MXE
RI
SE
E
Differential alternative splicing
&
Differential transcript usage
&
Statistically significant gene expression changes
59 events, 34 genes
Tier1 differential splicing events
Inclusion level difference
Gene
Event type
Number of differentially used transcripts associated with the event
ABCB7
AC093415.2
AMT
ANKHD1
AP5Z1
APBB3
ARIH1
BRD9
CRNDE
DLST
DPH5
DSN1
ERCC3
FOXRED1
GAS8
GUCY1B3
HERC2P9
HLTF
KIAA1033
METTL5
MFSD10
OXA1L
PROS1
PSTPIP1
RFNG
RPRD1A
SHKBP1
SNRPN
STAU2
TMEM214
TMEM218
TPM1
TYROBP
ZNF410

Figure 2.2. Integrative gene expression, alternative splicing, and transcript usage analyses categorize gene targets of mutant SF3B1. (A) PCA plots based on gene expression of the 3000 most highly variable genes color-coded by *SF3B1* mutation status and sign-coded by patient ID. (B) Heatmap showing distance of the indicated iPSC-HSPCs based on pairwise Pearson correlation of their gene expression profiles, color-coded by *SF3B1* mutation status and patient ID. (C) PCA plot based on inclusion levels of the differentially spliced events between $SF3B1^{K700E}$ and $SF3B1^{WT}$ iPSC-HSPCs. (D) Scatterplots comparing the mean inclusion levels of the differentially spliced events in $SF3B1^{K700E}$ vs $SF3B1^{WT}$ iPSC-HSPCs with different event types broken down by color, as indicated. (E) Schematic summarizing the integrative analysis used to derive a mutant SF3B1 signature of splicing events and genes and scatterplot showing the inclusion level difference of all 59 signature splicing events, corresponding to 34 genes. A positive y axis value indicates that the event is more frequently found in $SF3B1^{K700E}$ vs $SF3B1^{WT}$.

We observed that several of the transcripts used preferentially in $SF3B1^{K700E}$ vs $SF3B1^{WT}$ HSPCs were annotated as NMD (Figure 2.3). Notably, this increased use of NMD transcripts was also associated with decreased expression of the corresponding genes (*DLST*, *BRD9*, *KIAA1033*, *SHKBP1*, *GAS8*). This is consistent with previous findings showing that *SF3B1* mutations induce widespread use of abnormal cryptic 3' splice sites, leading to NMD of multiple transcripts[13,32].

The 59-splicing event signature is associated with *SF3B1* mutational status

To test whether the SF3B1 signature derived in iPSC-HSPCs is also found in primary patient samples, we interrogated transcriptome data from $CD34^+$ BM cells from 68 patients with MDS and 8 healthy individuals from a published dataset[9]. Thirty-one of the 59 tier 1 events (53%) were found differentially spliced (FDR < 0.05, |inclusion level difference| > 0.1) between *SF3B1*-mutated patients (SF3B1mut, n = 28) and patients with MDS without any SF mutations (SF-WT, n = 40). Twenty-eight of those were also found differentially spliced between *SF3B1*-mutated patients and healthy individuals (WT; n = 8; Figure 2.3; supplemental Figure 2.2D). This splicing signature was not found in events differentially spliced between MDS primary cells harboring other splicing factor mutations (*SRSF2*, *U2AF1*) and SF-WT MDS or healthy individuals and is thus specific to *SF3B1* mutations (supplemental Figure 2.2E). PCA based on the inclusion level of the mutant SF3B1 signature splicing events separated all samples (SF3B1mut; SF-WT; WT) based on *SF3B1* genotype, with the exception of 1 sample, annotated as SF-WT, which clustered together with the *SF3B1*-mutated samples (Figure 2.4). Examination of the RNA-seq data for sequence alterations in the *SF3B1* locus in this specific patient revealed a previously overlooked 6-bp in-frame deletion spanning the K700E hotspot (SF3B1p.K700_V701delKV; Figure 2.4). This demonstrates that the splicing signature derived in iPSC-HSPCs is also present in HSPCs of patients with MDS. Furthermore, patients with *SF3B1* mutations other than K700E clustered together with the $SF3B1^{K700E}$-mutated patients (Figure 2.4), which indicates that our signature is representative of a broader spectrum of *SF3B1* mutations.

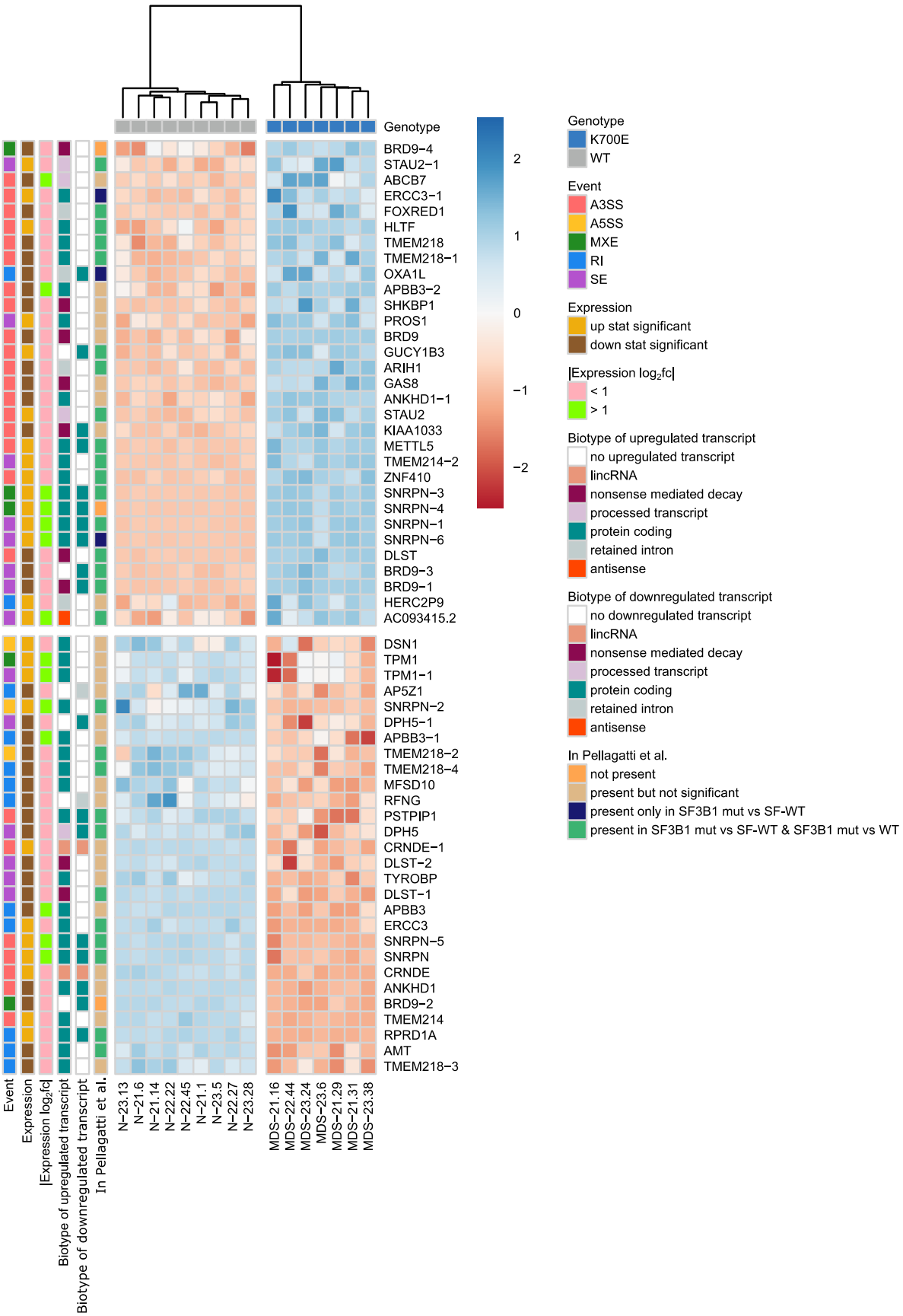

Genotype
BRD9−4
STAU2−1
ABCB7
ERCC3−1
FOXRED1
HLTF
TMEM218
TMEM218−1
OXA1L
APBB3−2
SHKBP1
PROS1
BRD9
GUCY1B3
ARIH1
GAS8
ANKHD1−1
STAU2
KIAA1033
METTL5
TMEM214−2
ZNF410
SNRPN−3
SNRPN−4
SNRPN−1
SNRPN−6
DLST
BRD9−3
BRD9−1
HERC2P9
AC093415.2
DSN1
TPM1
TPM1−1
AP5Z1
SNRPN−2
DPH5−1
APBB3−1
TMEM218−2
TMEM218−4
MFSD10
RFNG
PSTPIP1
DPH5
CRNDE−1
DLST−2
TYROBP
DLST−1
APBB3
ERCC3
SNRPN−5
SNRPN
CRNDE
ANKHD1
BRD9−2
TMEM214
RPRD1A
AMT
TMEM218−3
Event
Expression
|Expression log2fc|
Biotype of upregulated transcript
Biotype of downregulated transcript
In Pellagatti et al.
N−23.13
N−21.6
N−21.14
N−22.22
N−22.45
N−21.1
N−23.5
N−22.27
N−23.28
MDS−21.16
MDS−22.44
MDS−23.24
MDS−23.6
MDS−21.29
MDS−21.31
MDS−23.38
2
1
0
−1
−2
Genotype
K700E
WT
Event
A3SS
A5SS
MXE
RI
SE
Expression
up stat significant
down stat significant
|Expression log2fc|
< 1
> 1
Biotype of upregulated transcript
no upregulated transcript
lincRNA
nonsense mediated decay
processed transcript
protein coding
retained intron
antisense
Biotype of downregulated transcript
no downregulated transcript
lincRNA
nonsense mediated decay
processed transcript
protein coding
retained intron
antisense
In Pellagatti et al.
not present
present but not significant
present only in SF3B1 mut vs SF-WT
present in SF3B1 mut vs SF-WT & SF3B1 mut vs WT

Figure 2.3. Events of the mutant SF3B1 splicing signature. Heatmap showing the row normalized inclusion levels of the 59 signature events across HSPCs from all iPSC lines. For each row, color-coded side panels present metadata relevant to each event, including the log_2fc of expression of the respective genes, the biotypes of the up- and downregulated transcripts that are associated with the splicing events, and the presence of the events in the MDS patient dataset of Pellagatti et al[9], encoded as not present (signature events not present in any comparison); present but not significant (signature events that were not statistically significant or/and had an absolute inclusion level difference < 0.1 in both comparisons [SF3B1mut vs SF-WT and SF3B1mut vs WT, i.e., healthy individuals]); present only in SF3B1mut vs SF-WT (signature events statistically significant [FDR < 0.05] and with an absolute inclusion level difference > 0.1 only in the SF3B1mut vs SF-WT MDS patient comparison); and present in SF3B1mut vs SF-WT and SF3B1mut vs WT (signature events statistically significant [FDR < 0.05] and with an absolute inclusion level difference > 0.1 in both comparisons). The annotations of the transcript biotypes are derived from the Ensembl GRCh37 gtf annotation file. Each row represents one event labeled with the respective gene name followed by a number indicating distinct events.

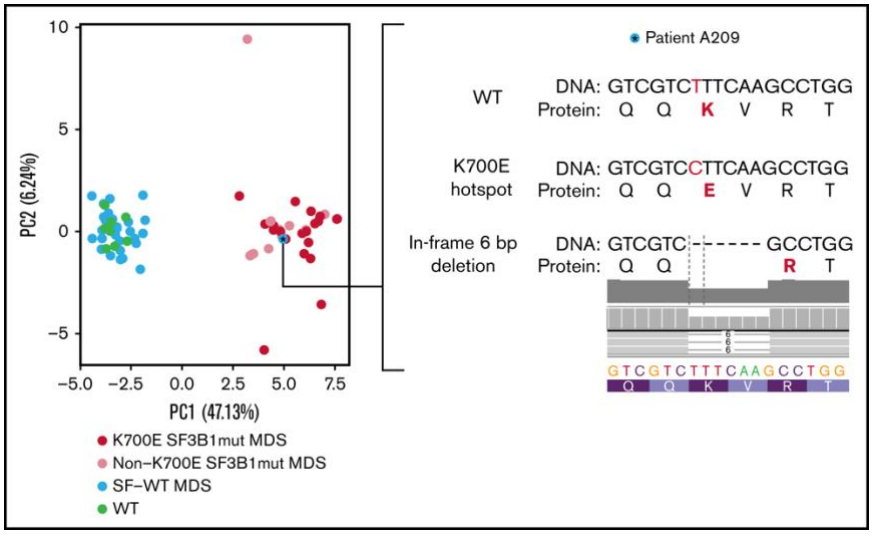


Figure 2.4. Splicing event signature separates *SF3B1*-mutated MDS cases. PCA plot based on the inclusion levels of the signature splicing events in the patient samples of Pellagatti et al[9], separating MDS *SF3B1*^{K700E}-mutated patients (K700E SF3B1mut MDS) and patients with *SF3B1* mutations other than K700E (non-K700E SF3B1mut MDS) from patients without SF mutations (SF-WT MDS) and healthy individuals (WT). The asterisk marks 1 patient annotated as SF-WT. Clustering of this sample together with the *SF3B1*-mutated cases prompted us to more closely interrogate the sequence of the *SF3B1* locus for any previously unidentified mutations. We thus discovered an in-frame 6-bp deletion (SF3B1p.K700_V701delKV) removing 2 amino acids, including the K700 hotspot.

Chromatin accessibility landscape of *SF3B1*^{K700E} HSPCs

To investigate the chromatin landscape of *SF3B1*^{K700E} cells, we performed ATAC sequencing (supplemental Methods) in sorted CD34$^+$/CD45$^+$ iPSC-HSPC samples paired to those used for RNA sequencing (3 *SF3B1*^{K700E} and 3 *SF3B1*^{WT} iPSC lines from each patient; supplemental Table 2.2) resulting in an ATAC-seq atlas of 56420 peaks. (Samples MDS-22.1, MDS-22.43, and N-21.1 did not pass quality control at the library preparation stage and were not included in the analyses.) PCA and hierarchical clustering based on chromatin accessibility grouped the iPSC lines by genotype (Figure 2.5A-B). Differential accessibility analysis revealed 3737

differentially accessible peaks between the *SF3B1*^{K700E} and *SF3B1*^{WT} HSPCs, 1527 of which were more accessible in the mutants (Figure 2.5C; supplemental Figure 2.3A). Differentially accessible peaks were predominantly localized in intronic and intergenic regions (supplemental Figure 2.3B). Chromatin accessibility changes correlated with gene expression changes in both directions (more accessible and upregulated; less accessible and downregulated; Figure 2.5D-E; supplemental Figure 2.3C). Next, we compared the chromatin accessibility profiles of the *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs to those defined in primary human hematopoietic cell types along the hematopoietic hierarchy[33] (supplemental Methods). Of the 56420 total ATAC-seq peaks called in the iPSC-HSPC dataset, 40568 overlapped with the peaks from Corces et al[33] (total = 98525). Differential accessibility analysis on these 40568 peaks resulted in 2757 differentially accessible peaks between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs. The pairwise Pearson correlation between read counts of these 2757 peaks in iPSC-HSPCs and the hematopoietic populations of Corces et al[33] showed that the chromatin landscapes of *SF3B1*^{K700E} cells resembled more those of megakaryocyte–erythroid progenitor cells and erythroid cells, whereas the chromatin landscape of *SF3B1*^{WT} cells resembled more that of granulocyte-monocyte progenitors and monocytes (Figure 2.5F). These results suggest a potential chromatin priming of *SF3B1*^{K700E} CD34^{+} HSPCs toward the erythroid rather than the myeloid lineage and may reflect the more prominent involvement of the erythroid lineage in the pathology and clinical presentation of MDS-RS.

Increased transcriptional activity of the TEAD family of transcription factors in *SF3B1*^{K700E} HSPCs

To identify transcriptional programs of potential importance to *SF3B1*^{K700E} HSPCs, we performed TF motif enrichment analysis in ATAC-seq peaks more accessible in *SF3B1*^{K700E} cells that were linked to genes upregulated in *SF3B1*^{K700E} cells (Figure 2.5E). This analysis revealed enrichment of motifs of several prototypical hematopoietic lineage TFs, such as those of the GATA, ETS, and AP-1 families. Unexpectedly, motifs of the TEAD family were also enriched (Figure 2.6A-C). Furthermore, regions more accessible and linked to upregulated genes in *SF3B1*^{K700E} cells that contained TEAD motifs overlapped with annotated TEAD binding sites (supplemental Figure 2.4A).

The TEAD family of TFs are best known as effectors of the Hippo signaling pathway, with important roles in various biological processes and malignancies, albeit no previous links to hematologic disease[34,35]. To further investigate a potential role for TEAD TFs in *SF3B1*^{K700E} HSPCs, we examined expression of the 4 members of the TEAD family *TEAD1-4* in *SF3B1*^{K700E} and *SF3B1*^{WT} cells. *TEAD2* and *TEAD4* were the TEAD family members expressed at the highest levels in both *SF3B1*^{K700E} and *SF3B1*^{WT} cells, including iPSC-HSPCs, as well as patient cells (Figure 2.6D; supplemental Figure 2.4B). All 4 *TEAD* genes were upregulated in the *SF3B1*^{K700E} compared with *SF3B1*^{WT} iPSC-HSPCs (Figure 2.6D).

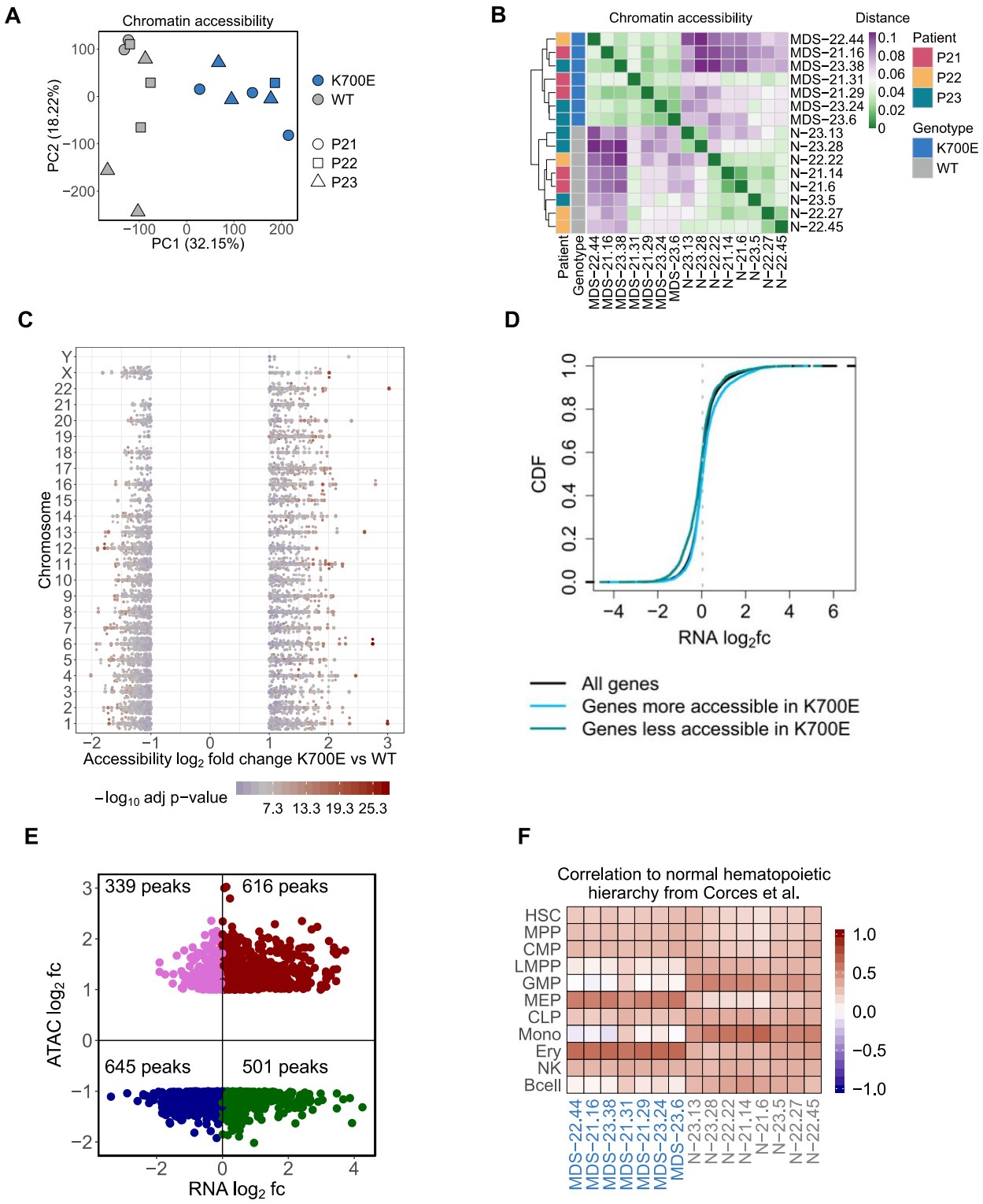


Figure 2.5. *SF3B1*^{K700E} HSPCs have altered chromatin landscapes. (A) PCA based on the accessibility of all peaks in the ATAC-seq atlas color-coded by *SF3B1* mutation status and sign-coded by patient ID. (B) Heatmap showing the distance of the HSPCs from the indicated iPSC lines, based on pairwise Pearson correlation of their chromatin accessibility landscapes, color-coded by *SF3B1* mutation status and patient ID. (C) Scatterplot showing the accessibility $\log_2$fc and

the adjusted *P* value of the differentially accessible peaks between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs per chromosome, color-coded by the adjusted *P* value. Each point represents a peak. (D) Cumulative distribution function (CDF) of the expression $\log_2$fc of genes more accessible in *SF3B1*^{K700E} HSPCs, genes less accessible in *SF3B1*^{K700E} HSPCs, and all genes, showing that genes more accessible in *SF3B1*^{K700E} HSPCs are upregulated (Kolmogorov–Smirnov [KS] test, *P* = 1.17e-07) and genes less accessible in *SF3B1*^{K700E} HSPCs are downregulated (KS test, *P* = 3.13e-16) compared with background. (E) Scatterplot showing the $\log_2$fc accessibility value of the differentially accessible peaks and the $\log_2$fc expression value of the linked gene (genes for which *P* value could not be calculated were excluded). (F) Heatmap showing Pearson correlation values of normalized read counts for ATAC-seq peaks that overlap between the indicated iPSC-HSPCs and primary normal hematopoietic cell subpopulations (hematopoietic stem cells [HSC], multipotent progenitors [MPP], common myeloid progenitors [CMP], lymphoid-primed multipotent progenitors [LMPP], granulocyte-monocyte progenitors [GMP], megakaryocyte-erythrocyte progenitors [MEP], common lymphoid progenitors [CLP], monocytes [Mono], erythroid cells [Ery], natural killer cells [NK], and B cells) from Corces et al.[33].

To experimentally test whether TEAD transcriptional activity is higher in *SF3B1*^{K700E} cells, we transduced *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs with a luciferase construct (implemented at the Papapetrou laboratory and its experimental process is not part of this thesis) reporting TEAD activity. Reporter activity was higher or trended higher in *SF3B1*^{K700E} compared with *SF3B1*^{WT} iPSC-HSPCs from 2 of the 3 patients (Figure 2.6E). TEAD is best known as an effector of the Hippo signaling pathway and is bound to DNA as a complex with YAP or TAZ transcriptional coactivators[36]. To test the activity of the Hippo pathway in our cells, we performed immunoblots (implemented at the Papapetrou laboratory and the experimental process is not part of this thesis) in *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs from 2 of the patients. Although we confirmed TEAD expression at the protein level, we did not detect YAP activation (phosphorylated form pYAP^{S127}) or expression of YAP or TAZ (supplemental Figure 2.4C). These results, collectively, support a Hippo pathway-independent increase of TEAD expression and transcriptional activity in *SF3B1*^{K700E} HSPCs.

2.5. Discussion

Previous studies have shown that iPSC models of myeloid malignancies capture molecular characteristics of the disease and can be used to discover new mechanisms and therapeutic vulnerabilities, further corroborated by the present study[37–40]. Hereby, we harnessed sequencing data from patient-derived genetically matched WT and *SF3B1*-mutated induced pluripotent stem cell (iPSC) lines. The genetically matched conditions and the availability of biological replicates to control for any effects of the reprogramming process (nongenetic line-to-line variability upon cell line generation) and of the patient's genetic background on the transcriptome, were critical components of this study.

Our study was powered by a multi-stage data fusion framework with which we were able to assess the combination of the effects of the *SF3B1*^{K700E} mutation across parallel levels of deregulation of the transcriptome toward deriving a tier-based classification of splicing events. Specifically, this framework systematically evaluates the effects *SF3B1*^{K700E} mutation on splicing, transcript usage and gene expression, merges signals from the data views representative of these 3 processes, and leads to a comprehensive *SF3B1*^{K700E} splicing signature. These integrated analyses validated several known gene candidates, such as *ANKHD1[9]*, *METTL5* (A3SS event)[9,13], *ABCB7* (A3SS event)[11,41,42] and *BRD9* (SE event)[13].

Additionally, the genes *DPH5*, *COASY*, *ZDHHC16*, *TMEM214*, and *EI24*, previously cataloged as mis-spliced in *SF3B1* mutant cells, are also included in our tier 2 and tier 1 set[13]. Furthermore, we nominate several new splicing events in genes not previously reported mis-spliced by mutant SF3B1 that warrant further investigation for their relevance to the pathogenesis of MDS. The diversity of mis-splicing events, many of which are found across different models of *SF3B1*^{K700E} mutation, may suggest a multifactorial disease pathogenesis. In addition, the specificity of the mutant SF3B1 signature, derived from the iPSC lines and validated in primary patient samples, identifies atypical mutations involving the K700 hotspot, such as the SF3B1p.K700_V701delKV that we report here, as functionally equivalent to the K700E mutation, and can thus be further used to evaluate the role of putative pathogenic variants in *SF3B1[43]*.

Our study is the first to characterize the chromatin landscape of *SF3B1*^{K700E} HSPCs. Interestingly, we report potential "priming" at the chromatin level of *SF3B1*^{K700E} HSPCs toward the erythroid over the myeloid lineage, a finding that may be related to the preferential involvement of the erythroid lineage in MDS and, in particular, MDS-RS. It is unclear whether any of the global chromatin accessibility changes that we report here are a direct consequence of missplicing (for example, of a chromatin regulator gene, such as *BRD9*[13], or a pioneer transcription factor). Likely, at least some of them reflect differences in differentiation state and lineage priming as an indirect consequence of the *SF3B1*^{K700E} mutation. Because reprogramming to pluripotency (upon the experimental process) effectively erases the epigenome of the somatic cell, differences found between mutant and WT cells across replicates can be solely attributed to genotype.

Several master hematopoietic lineage TF motifs were present in chromatin regions that were differentially accessible between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs, which may underlie the differentiation and colony formation impairment of these cells. Interestingly, our chromatin accessibility analyses, followed by functional studies, lend support to a putative role for the TEAD TFs in the context of *SF3B1*^{K700E} mutation. The relevance of elevated TEAD activity to the pathogenesis of MDS-RS and its link to *SF3B1* mutations will need to be validated in further studies involving assessment of TEAD binding to DNA and functional experiments, such as genetic perturbation of TEAD factors, in our iPSC models, as well as potential validation of the findings in primary patient cells. Pending further investigation, this novel finding may point to a new disease mechanism and possible therapeutic vulnerabilities specific to *SF3B1*^{K700E} cells.

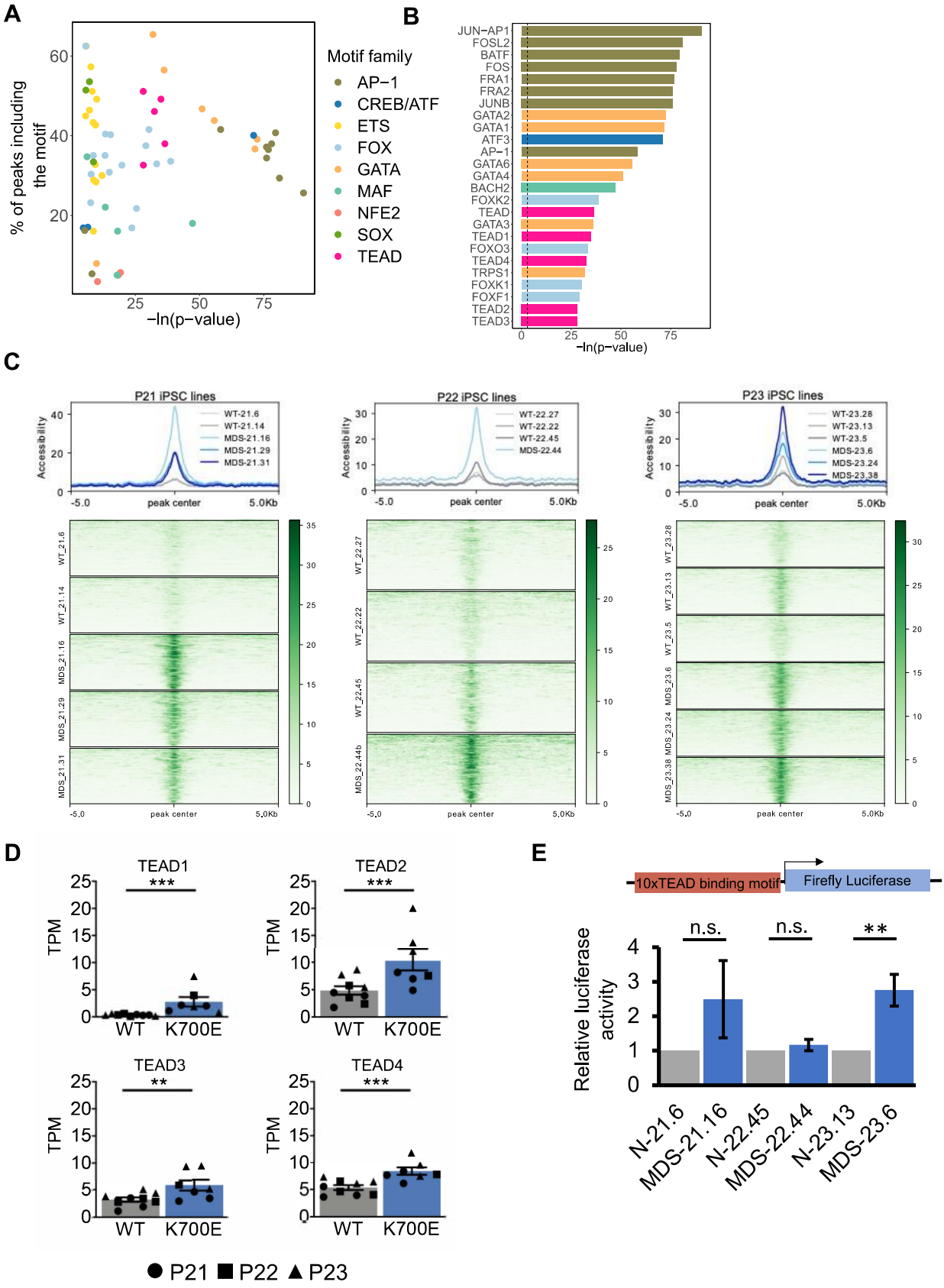

A
% of peaks including the motif
−ln(p−value)
Motif family
AP−1
CREB/ATF
ETS
FOX
GATA
MAF
NFE2
SOX
TEAD
B
JUN−AP1
FOSL2
BATF
FOS
FRA1
FRA2
JUNB
GATA2
GATA1
ATF3
AP−1
GATA6
GATA4
BACH2
FOXK2
TEAD
GATA3
TEAD1
FOXO3
TEAD4
TRPS1
FOXK1
FOXF1
TEAD2
TEAD3
−ln(p−value)
C
P21 iPSC lines
P22 iPSC lines
P23 iPSC lines
Accessibility
peak center
5.0Kb
WT-21.6
WT-21.14
MDS-21.16
MDS-21.29
MDS-21.31
WT-22.27
WT-22.22
WT-22.45
MDS-22.44
WT-23.28
WT-23.13
WT-23.5
MDS-23.6
MDS-23.24
MDS-23.38
D
TEAD1
TEAD2
TEAD3
TEAD4
TPM
WT
K700E
● P21 ■ P22 ▲ P23
E
10xTEAD binding motif
Firefly Luciferase
Relative luciferase activity
n.s.
n.s.
**
N-21.6
MDS-21.16
N-22.45
MDS-22.44
N-23.13
MDS-23.6

Figure 2.6. Increased transcriptional activity of TEAD TFs in *SF3B1*^{K700E} HSPCs. (A) TF motifs enriched in peaks more accessible in *SF3B1*^{K700E} compared with *SF3B1*^{WT} HSPCs and linked to upregulated genes. (B) Most statistically significant TF motifs enriched in peaks more accessible in *SF3B1*^{K700E} compared with *SF3B1*^{WT} HSPCs and linked to upregulated genes. (C) Tornado plots showing the normalized accessibility signal in peaks more accessible in *SF3B1*^{K700E} compared with *SF3B1*^{WT} HSPCs and linked to upregulated genes that contain TEAD motifs. (D) Expression levels of TEAD family genes in iPSC-HSPCs. Mean and SEM of transcripts per million (TPM) values from RNA-seq are shown. ***P*adj ≤ .01; ****P*adj ≤ .001. (E) TEAD reporter activity in HSPCs from the indicated iPSC lines. Mean and SEM of 2 to 5 independent differentiation and transduction experiments per line are shown. n.s., not significant; ***P* ≤ .01.

2.6. Supplementary

2.6.1. Supplemental methods

Targeted gene sequencing

Variant calling and annotation, filtering for artifacts and copy number identification was performed as previously described[3,15].

GSEA analysis

GSEA analysis was performed on the differentially expressed genes between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs using the R-package clusterProfiler[44]. Only gene sets with Benjamini-Hochberg (BH) adjusted p-value < 0.05 were considered.

Computation of statistical significance in differential transcript usage

The statistical significance of the change in transcript usage between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs was assessed in a two stage process using the stageR R-package. The first stage (screening phase) identified the genes with evidence of differential transcript usage (DTU) and the second stage (confirmation phase) identified the transcripts within these genes that participate in the DTU.

Integration Framework

As shown in supplemental Figure 2.2A, the categorization of differential splicing events to 5 tier-based classes (tier 1, tier 2, tier 3, tier 4 and tier 5) depends on the following criteria: (1) Presence of at least one differentially used transcript paired to the event (2) Statistically significant expression $\log_2$fc of the respective gene (3) Statistical significance of the event (FDR value among the lowest 20 FDR values across all events). Based on the criteria met, the events are rewarded with scores (score of 2 for meeting criterion 1 and score of 1 for meeting each of the rest). According to their total, events are classified to a specific tier as shown in the schema of supplemental Figure 2.2A. Events that meet all criteria (total score of 4) comprise the tier 1 (signature) class.

ATAC sequencing

50,000 MACS-sorted $CD34^+/CD45^+$ cells from each individual iPSC line were processed as follows: nuclei were isolated by lysis with 50 ul of ATAC lysis buffer (10 mM Tris pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin) and washed with 1 mL of ATAC wash buffer (10 mM Tris pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Tween-20). Cell lysates were spun to obtain nuclear pellets, which were subjected to transposase reaction using the Illumina Nextera DNA Sample Preparation Kit according to the manufacturer's instructions.

Peak Calling, assignment of ATAC-seq peaks to genes, differential accessibility of genes & motif enrichment analysis

ATAC-seq peaks for each replicate of the cohort were called using MACS2[30] with a low pass whole genome sequencing background sample and the following parameters (--min-length 100 -f BAMPE). Each ATAC-seq peak was assigned to the gene with the closest Transcription Start Site (TSS) using HOMER[45] with the Ensembl GRCh37 gtf annotation file. To determine accessibility changes at the gene level, we considered the regulation of the surrounding peak regions as well as the proximity of each gene to a differentially accessible peak. Specifically, a gene was regarded as differentially accessible if: (1) There was at least 1 differentially accessible peak within 50 kb of its TSS; or (2) The distribution of accessibility $\log_2$fc of all peaks within 50 kb upstream and downstream of its TSS was significantly shifted (Benjamini-Hochberg [BH] adjusted p-value < 0.05), as compared to the ATAC-seq atlas background distribution by a Kolmogorov-Smirnov (KS) test. Motif enrichment analysis was performed with HOMER using known motifs in the HOMER default database.

Correlation of chromatin accessibility to normal hematopoiesis

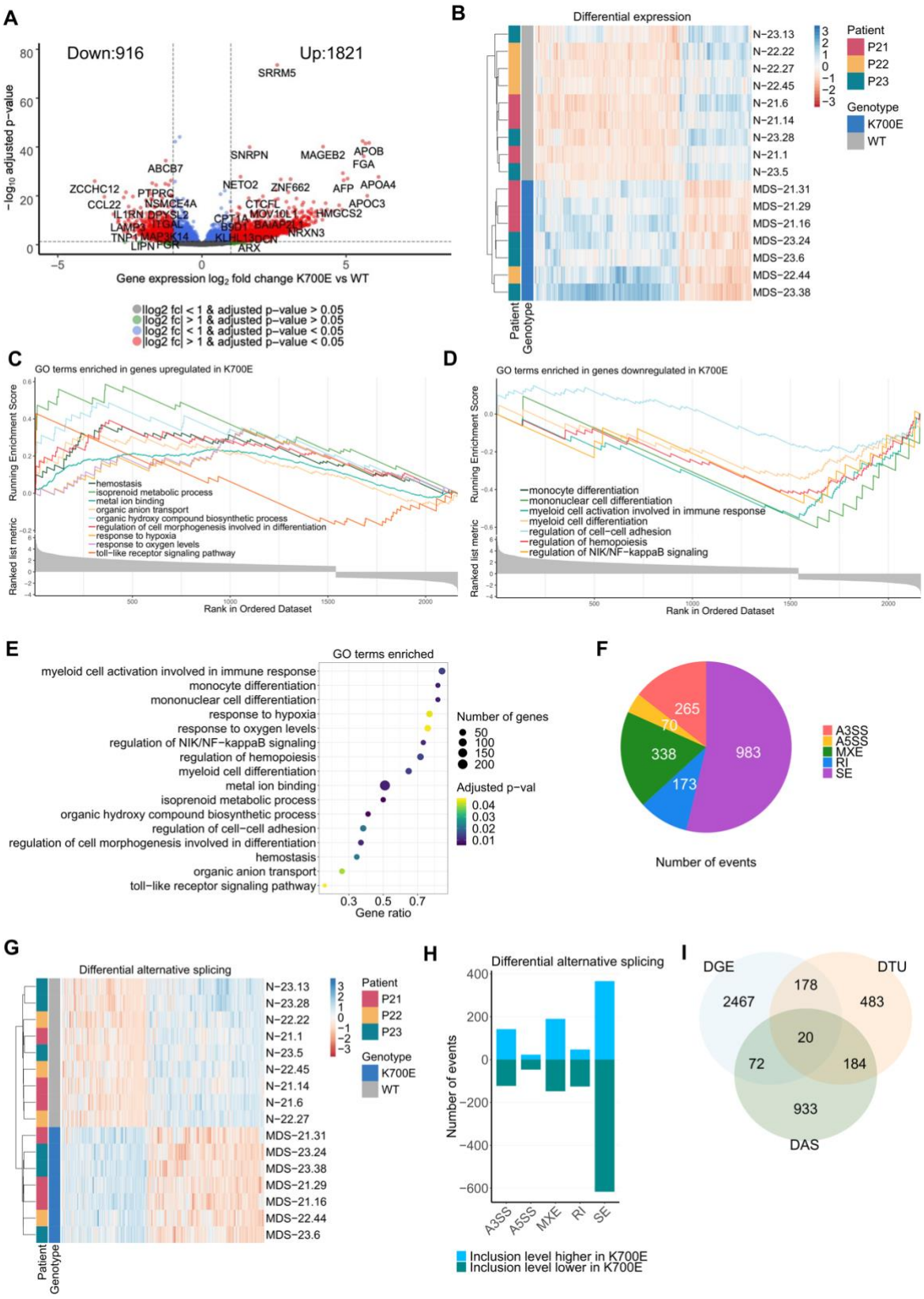
To compare the accessibility profiles of our iPSC-HSPCs to normal primary hematopoietic populations, we obtained raw bulk ATAC-seq data of 67 samples (7 HSC, 6 MPP, 3 LMPP, 8 CMP, 7 GMP, 7 MEP, 6 Mono, 8 Ery, 5 CLP, 6 NK, 4 B cell) from a published dataset[33]. These were processed as described above for the iPSC-HSPC data. Pairwise Pearson correlation between iPSC-HSPCs and the normal hematopoietic populations was computed based on a set of differentially accessible peaks between the $SF3B1^{K700E}$ and $SF3B1^{WT}$ iPSC-HSPCs that overlapped with the ATAC-seq atlas of the Corces et al. samples

Data and code availability

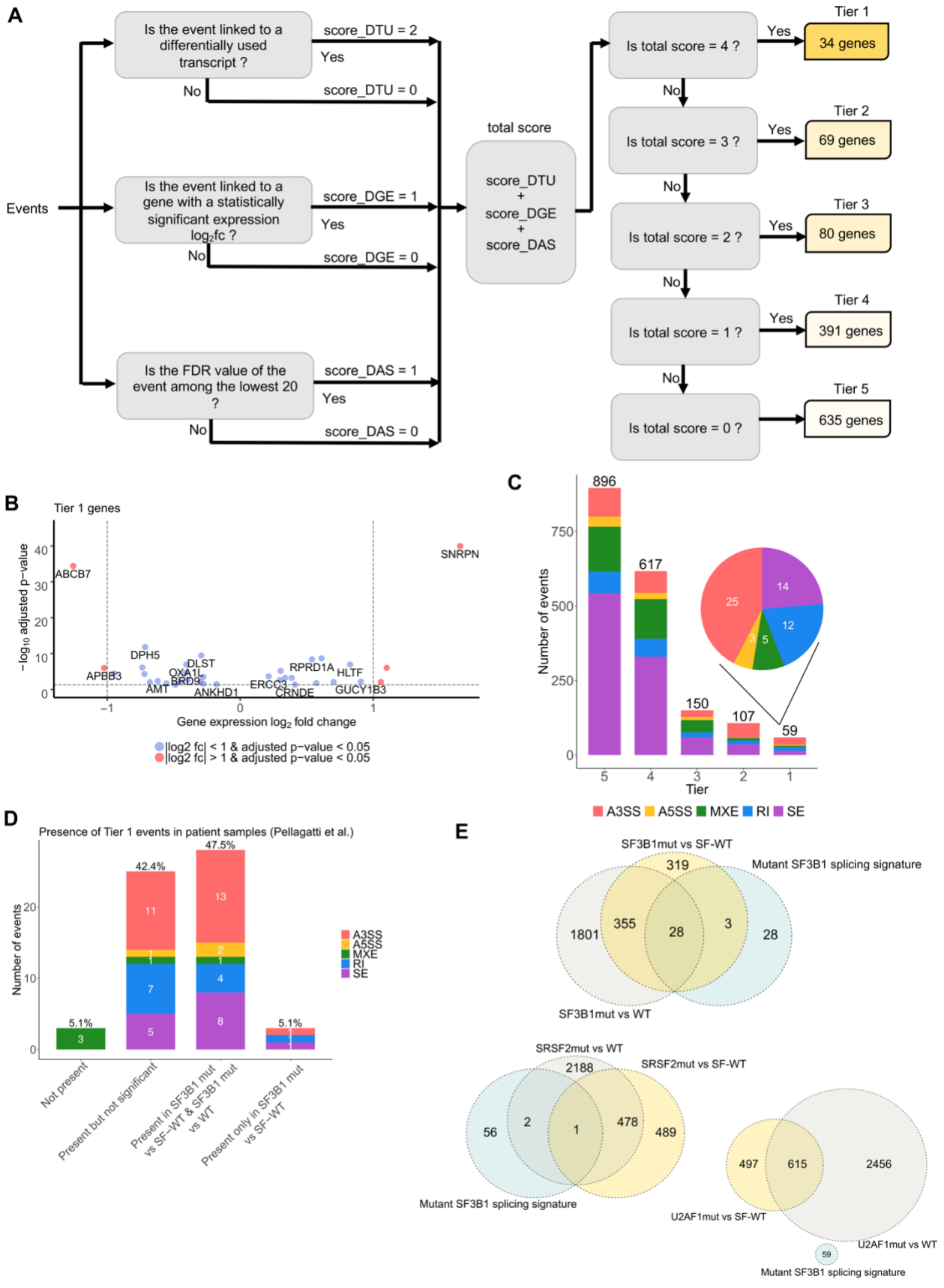
The work presented in this chapter has been published here (Blood Advances journal).

Data preprocessing and analysis was conducted using R 3.5.2 and bash scripting. A github repository containing the code used in generating the figures and the analysis results is available at the Papaemmanuil lab github page (https://github.com/papaemmelab/MDS_SF3B1_iPSC). The data used for this project are deposited in the Gene Expression Omnibus (GEO) with accession number GSE184246.

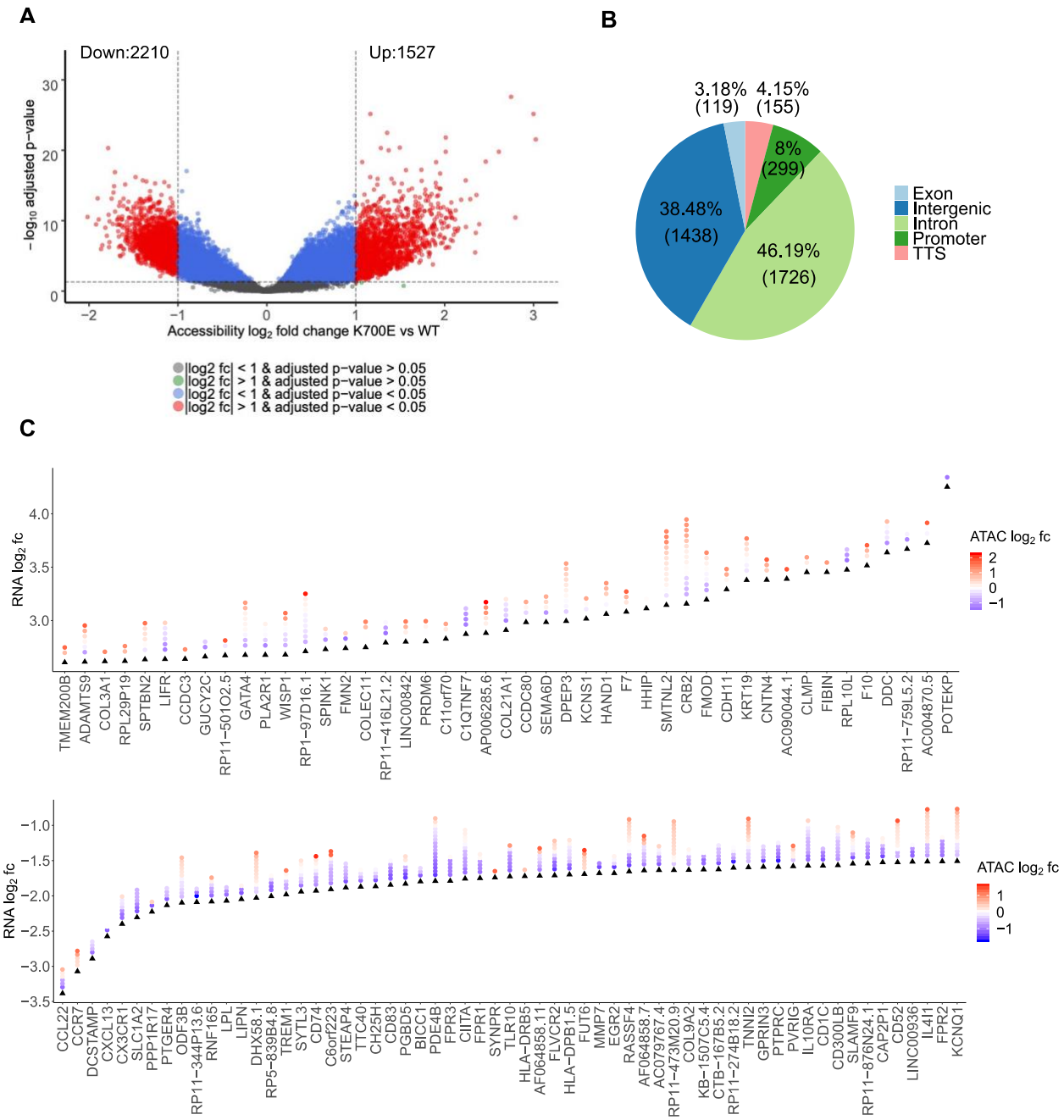
2.6.2. Supplemental figures



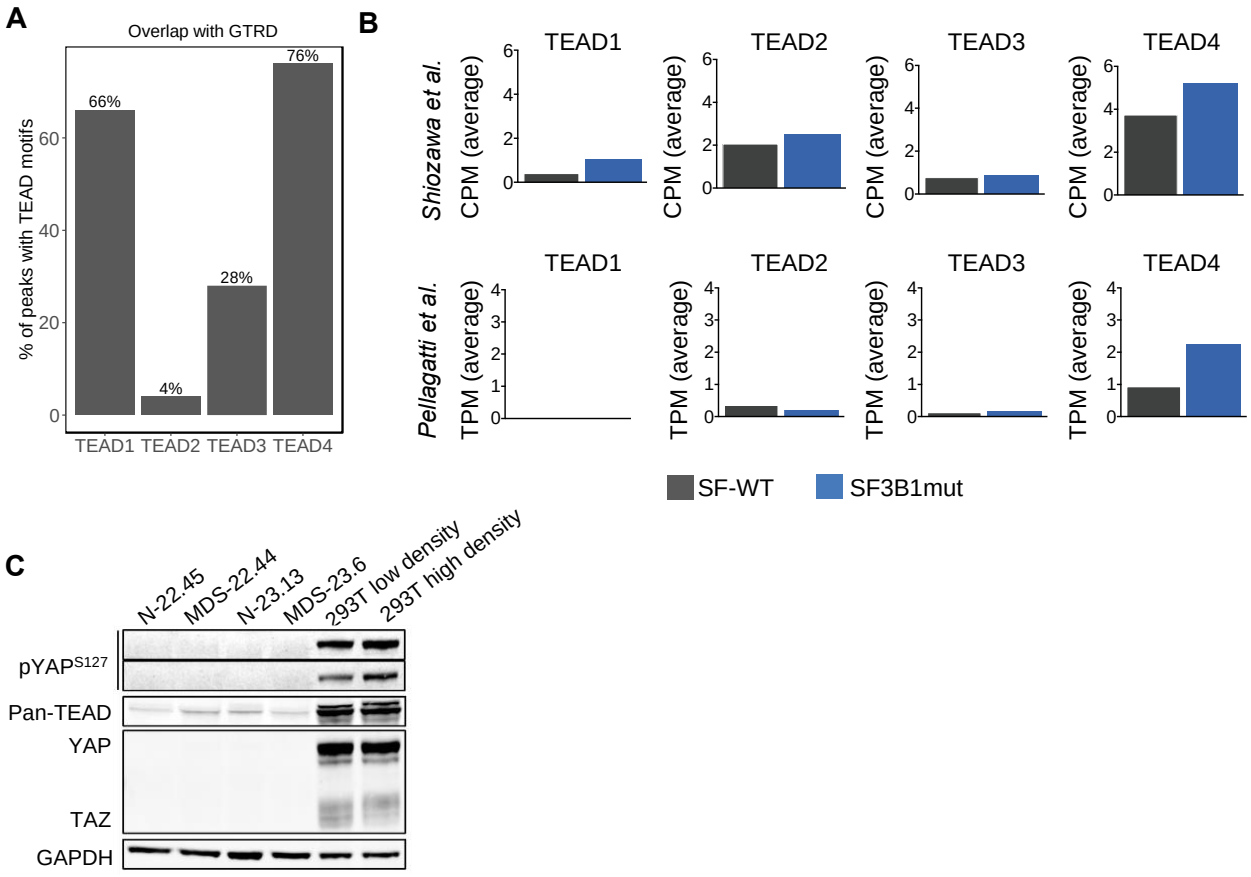
Supplemental Figure 2.1. Gene expression, splicing and transcript usage analyses. (A) Volcano plot of differentially expressed genes between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs. (B) Column-normalized heatmap of gene expression values of the differentially expressed genes between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs, color-coded by *SF3B1* mutation status and patient ID. (C,D) GSEA plots of gene ontology (GO) gene sets enriched in genes upregulated (C) or downregulated (D) in *SF3B1*^{K700E} vs *SF3B1*^{WT} iPSC-HSPCs. (E) Scatterplot of enriched GO terms. (F) Distribution of the differential splicing events between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs by event type. (G) Column-normalized heatmap of inclusion levels of the differentially spliced events between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs, color-coded by *SF3B1* mutation status and patient ID. (H) Differential splicing events between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs. (I) Overlap between differentially expressed genes, genes with differentially used transcripts and genes linked to differential splicing events between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs.


A
Events
Is the event linked to a differentially used transcript ?
score_DTU = 2
Yes
No
score_DTU = 0
Is the event linked to a gene with a statistically significant expression log2fc ?
score_DGE = 1
Yes
No
score_DGE = 0
Is the FDR value of the event among the lowest 20 ?
score_DAS = 1
Yes
No
score_DAS = 0
total score
score_DTU + score_DGE + score_DAS
Is total score = 4 ?
Yes
Tier 1
34 genes
No
Is total score = 3 ?
Yes
Tier 2
69 genes
No
Is total score = 2 ?
Yes
Tier 3
80 genes
No
Is total score = 1 ?
Yes
Tier 4
391 genes
No
Is total score = 0 ?
Tier 5
635 genes
B
Tier 1 genes
−log10 adjusted p−value
Gene expression log2 fold change
ABCB7
SNRPN
APBB3
DPH5
DLST
OXA1L
BRD9
AMT
ANKHD1
RPRD1A
HLTF
ERCC3
CRNDE
GUCY1B3
|log2 fc| < 1 & adjusted p−value < 0.05
|log2 fc| > 1 & adjusted p−value < 0.05
C
Number of events
Tier
896
617
150
107
59
25
14
3
5
12
A3SS A5SS MXE RI SE
D
Presence of Tier 1 events in patient samples (Pellagatti et al.)
Number of events
42.4%
47.5%
5.1%
5.1%
Not present
Present but not significant
Present in SF3B1 mut vs SF−WT & SF3B1 mut vs WT
Present only in SF3B1 mut vs SF−WT
A3SS
A5SS
MXE
RI
SE
E
SF3B1mut vs SF-WT
Mutant SF3B1 splicing signature
SF3B1mut vs WT
1801
355
28
3
28
319
SRSF2mut vs WT
SRSF2mut vs SF-WT
Mutant SF3B1 splicing signature
56
2
1
478
489
2188
U2AF1mut vs SF-WT
U2AF1mut vs WT
Mutant SF3B1 splicing signature
497
615
2456
59

Supplemental Figure 2.2. Integration Framework (A) Schema used for the categorization of differential splicing events to 5 tier-based classes: tier 1, tier 2, tier 3, tier 4 and tier 5. The classification of each event to one of these classes is based on the following: (1) Presence of at least one differentially used transcript paired to the event (2) Statistically significant expression log_2fc of the respective gene (3) Statistical significance of the event (FDR value among the lowest 20 FDR values across all events). (B) Volcano plot of the tier 1 genes. (C) Differential splicing events in each tier-based class color-coded per event type. (D) Intersection of the tier 1 splicing events with the Pellagatti et al.[9] dataset. Not present: tier 1 events not present in any comparison. Present but not significant: tier 1 events that were not statistically significant or/and had an absolute inclusion level difference < 0.1 in both comparisons (SF3B1mut vs SF-WT & SF3B1mut vs WT, i.e. healthy individuals). Present in SF3B1mut vs SFWT & SF3B1mut vs WT: tier 1 events statistically significant (FDR < 0.05) and with an absolute inclusion level difference > 0.1 in both comparisons. Present only in SF3B1mut vs SF-WT: tier 1 events statistically significant (FDR < 0.05) and with an absolute inclusion level difference > 0.1 only in the SF3B1mut vs SF-WT MDS patient comparison. (E) Venn diagrams showing overlap of differential splicing events present in the respective comparisons from MDS patients from the Pellagatti et al.[9] dataset and our mutant SF3B1 splicing signature. There is minimal or no overlap between the mutant SF3B1 splicing signature and differential splicing events found in MDS cases with *SRSF2* (lower left) or *U2AF1* (lower right) mutations.



Supplemental Figure 2.3. Chromatin accessibility analyses. (A) Volcano plot showing the differentially accessible peaks between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs (B) Genomic distribution of the differentially accessible peaks between *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC-HSPCs. (TTS: transcription termination site) (C) Diamond plot showing the differentially accessible genes with the highest and lowest expression log_2fc, together with the accessibility log_2fc of the peaks associated with these respective genes. The black triangle shows the expression log2fc of each gene, and the points above correspond to all peaks associated with each gene. The peaks are color-coded based on their accessibility log_2fc.



Supplemental Figure 2.4. TEAD transcriptional activity in *SF3B1*^{K700E} and *SF3B1*^{WT} iPSC HSPCs. (A) Percentage of peaks more accessible in *SF3B1*^{K700E} compared to *SF3B1*^{WT} HSPCs and linked to upregulated genes containing TEAD motifs that overlap with TEAD binding sites from the Gene Transcription Regulation Database (GTRD) across all samples. (B) Expression of TEAD family genes in primary patient samples from Shiozawa et al.[11] (upper panels) and Pellagatti et al.[9] (lower panels). MDS SF3B1mut: MDS with isolated *SF3B1* mutation for the Shiozawa et al.[11] dataset; MDS with isolated *SF3B1 K700E* mutation for the Pellagatti et al.[9] dataset., MDS SF-WT: MDS with no mutations in *SF3B1*, *SRSF2*, *U2AF1*, *ZRSR2*, *U2AF2*, *PRPF8*, *SF1* genes for the Shiozawa et al.[11] dataset; MDS with no mutations in *SF3B1*, *SRSF2*, *U2AF1*, *ZRSR2* genes for the Pellagatti et al.[9] dataset. CMP: counts per million; TPM: transcripts per million. (C) Phosphorylated YAP, TEAD, total YAP and TAZ expression in HSPCs from the indicated iPSC lines on day 12 of differentiation and in 293T cells as positive control.

2.6.3. Supplemental tables

Supplemental Table 2.1. Clinical, cytogenetic and mutational profile of MDS-RS patients selected for this study. A pericentric inversion of chromosome 9 found in patient 23 is a polymorphic chromosomal rearrangement not linked to MDS[46,47]. The patients did not harbor additional MDS/AML-driver mutations other than the *SF3B1 K700E*. (They were selected for the study on the basis of isolated *SF3B1 K700E* mutation.)

Patient	Sex	Age	Diagnosis	*SF3B1 K700E* VAF	Cytogenetics
P21	Male	74	MDS-RS	0.42	46, XY
P22	Female	65	MDS-RS	0.37	46, XX, +mar
P23	Male	84	MDS-RS	0.37	46,XY, inv(9)(p11q13)

Supplemental Table 2.2. All iPSC lines used in this study.

Patient	iPSC line	Genotype	RNA-Seq	ATAC-Seq
P21	N-21.1	WT/WT	+	-
P21	N-21.6	WT/WT	+	+
P21	N-21.14	WT/WT	+	+
P21	MDS-21.16	WT/K700E	+	+
P21	MDS-21.29	WT/K700E	+	+
P21	MDS-21.31	WT/K700E	+	+
P22	N-22.22	WT/WT	+	+
P22	N-22.27	WT/WT	+	+
P22	N-22.45	WT/WT	+	+
P22	MDS-22.1	WT/K700E	-	-
P22	MDS-22.43	WT/K700E	-	-
P22	MDS-22.44	WT/K700E	+	+
P23	N-23.5	WT/WT	+	+
P23	N-23.13	WT/WT	+	+
P23	N-23.28	WT/WT	+	+
P23	MDS-23.6	WT/K700E	+	+
P23	MDS-23.24	WT/K700E	+	+
P23	MDS-23.38	WT/K700E	+	+

Supplemental Table 2.3. Tier-based classification of events and qualitative levels of evidence.

Tier	Qualitative levels of evidence
1	DTU ∩ top ranked in splicing ∩ gene shows statistically significant expression change
2	DTU ∩ top ranked in splicing or DTU ∩ gene shows statistically significant expression change
3	DTU or top ranked in splicing ∩ gene shows statistically significant expression change
4	top ranked in splicing or gene shows statistically significant expression change
5	any other differential splicing event

Supplemental Table 2.4. Mutant SF3B1 splicing signature.

Event Type	Gene	Chr	Starting position	Ending Position	FDR	Inclusion level difference	Label	Expression	|Expression_log2fc|
A3SS	ABCB7	X	74291343	74291539	0	0.127	ABCB7	down stat significant	>1
A3SS	ANKHD1	5	139818045	139818202	0	-0.524	ANKHD1	down stat significant	<1
A3SS	ANKHD1	5	139818078	139818202	0	0.458	ANKHD1-1	down stat significant	<1
A3SS	APBB3	5	139941171	139941307	0	0.321	APBB3-2	down stat significant	>1
A3SS	ARIH1	15	72862504	72862648	0	0.119	ARIH1	down stat significant	<1
A3SS	BRD9	5	869359	869519	0	0.442	BRD9	down stat significant	<1
A3SS	CRNDE	16	54954209	54954322	0	-0.651	CRNDE	up stat significant	<1
A3SS	CRNDE	16	54954209	54954322	0	-0.372	CRNDE-1	up stat significant	<1
A3SS	DLST	14	75356580	75356655	0	0.389	DLST	down stat significant	<1
A3SS	ERCC3	2	128046912	128047095	0	0.147	ERCC3-1	up stat significant	<1
A3SS	FOXRED1	11	126143210	126143349	0	0.114	FOXRED1	down stat significant	<1
A3SS	GAS8	16	90097583	90097904	0	0.461	GAS8	down stat significant	<1

A3SS	GUCY1B3	4	156723364	156723731	0	0.109	GUCY1B3	up stat significant	<1
A3SS	HLTF	3	148759276	148759467	0	0.246	HLTF	up stat significant	<1
A3SS	KIAA1033	12	105514866	105514982	0	0.28	KIAA1033	down stat significant	<1
A3SS	METTL5	2	170668966	170669034	0	0.141	METTL5	up stat significant	<1
A3SS	PSTPIP1	15	77328142	77328276	0	-0.217	PSTPIP1	down stat significant	<1
A3SS	SHKBP1	19	41084353	41084448	0	0.108	SHKBP1	down stat significant	<1
A3SS	SNRPN	15	25219434	25219603	0	-0.296	SNRPN-5	up stat significant	>1
A3SS	SNRPN	15	25219457	25219603	0	-0.145	SNRPN	up stat significant	>1
A3SS	STAU2	8	74621266	74621412	0	0.315	STAU2	up stat significant	<1
A3SS	TMEM214	2	27260657	27260760	0	-0.471	TMEM214	up stat significant	<1
A3SS	TMEM218	11	124972027	124972247	0	0.531	TMEM218	down stat significant	<1
A3SS	TMEM218	11	124972027	124972247	0	0.355	TMEM218-1	down stat significant	<1
A3SS	ZNF410	14	74360478	74360635	0	0.296	ZNF410	up stat significant	<1
A5SS	DSN1	20	35399275	35399876	2.97 E-12	-0.111	DSN1	up stat significant	<1
A5SS	SNRPN	15	25212175	25212387	0	-0.14	SNRPN-2	up stat significant	>1
A5SS	TMEM218	11	124972532	124972705	4.60 E-08	-0.188	TMEM218-2	down stat significant	<1
MXE	BRD9	5	868721	869234	0	0.193	BRD9-4	down stat significant	<1
MXE	BRD9	5	869359	869509	0	-0.197	BRD9-2	down stat significant	<1
MXE	SNRPN	15	25212175	25212299	0	0.192	SNRPN-3	up stat significant	>1
MXE	SNRPN	15	25212175	25212387	0	0.153	SNRPN-4	up stat significant	>1
MXE	TPM1	15	63353396	63353472	0	-0.101	TPM1	up stat significant	>1
RI	AMT	3	49454210	49455151	0	-0.283	AMT	down stat significant	<1
RI	AP5Z1	7	4829462	4830222	1.41 E-10	-0.293	AP5Z1	down stat significant	<1

RI	APBB3	5	139941171	139941434	0	-0.655	APBB3	down stat significant	>1
RI	APBB3	5	139941171	139941812	0	-0.278	APBB3-1	down stat significant	>1
RI	ERCC3	2	128046912	128047400	0	-0.329	ERCC3	up stat significant	<1
RI	HERC2P9	15	28881632	28882253	4.07 E-10	0.401	HERC2P9	up stat significant	<1
RI	MFSD10	4	2934326	2934936	7.21 E-13	-0.115	MFSD10	down stat significant	<1
RI	OXA1L	14	23239401	23239834	0	0.199	OXA1L	down stat significant	<1
RI	RFNG	17	80007552	80007882	1.01 E-13	-0.171	RFNG	down stat significant	<1
RI	RPRD1A	18	33605560	33607038	0	-0.822	RPRD1A	up stat significant	<1
RI	TMEM218	11	124972027	124972705	0	-0.444	TMEM218-3	down stat significant	<1
RI	TMEM218	11	124972027	124972705	0	-0.312	TMEM218-4	down stat significant	<1
SE	AC093415.2	3	37892914	37892983	1.22 E-12	0.302	AC093415.2	up stat significant	>1
SE	BRD9	5	869359	869509	0	0.244	BRD9-3	down stat significant	<1
SE	BRD9	5	869359	869519	0	0.359	BRD9-1	down stat significant	<1
SE	DLST	14	75349293	75349327	0	-0.324	DLST-1	down stat significant	<1
SE	DLST	14	75352288	75352337	0	-0.114	DLST-2	down stat significant	<1
SE	DPH5	1	101458192	101458296	0	-0.131	DPH5	down stat significant	<1
SE	DPH5	1	101490864	101491022	0	-0.143	DPH5-1	down stat significant	<1
SE	PROS1	3	93647545	93647641	0	0.38	PROS1	up stat significant	<1
SE	SNRPN	15	25212175	25212299	0	0.283	SNRPN-1	up stat significant	>1
SE	SNRPN	15	25212175	25212387	0	0.155	SNRPN-6	up stat significant	>1
SE	STAU2	8	74621266	74621412	0	0.148	STAU2-1	up stat significant	<1
SE	TMEM214	2	27260682	27260760	0	0.457	TMEM214-2	up stat significant	<1
SE	TPM1	15	63353396	63353472	0	-0.167	TPM1-1	up stat significant	>1
SE	TYROBP	19	36398631	36398664	0	-0.149	TYROBP	down stat significant	<1

2.7. References

1. Cazzola M. Myelodysplastic Syndromes. N Engl J Med. 2020;383: 1358–1374.

2. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011;478: 64–69.

3. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood. 2013;122: 3616–27; quiz 3699.

4. Haferlach T, Nagata Y, Grossmann V, Okuno Y, Bacher U, Nagae G, et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia. 2014;28: 241–247.

5. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med. 2011;365: 1384–1395.

6. Malcovati L, Ambaglio I, Elena C. The genomic landscape of myeloid neoplasms with myelodysplasia and its clinical implications. Curr Opin Oncol. 2015;27: 551–559.

7. Malcovati L, Stevenson K, Papaemmanuil E, Neuberg D, Bejar R, Boultwood J, et al. SF3B1-mutant MDS as a distinct disease subtype: a proposal from the International Working Group for the Prognosis of MDS. Blood. 2020;136: 157–170.

8. Matera AG, Wang Z. A day in the life of the spliceosome. Nat Rev Mol Cell Biol. 2014;15: 108–121.

9. Pellagatti A, Armstrong RN, Steeples V, Sharma E, Repapi E, Singh S, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood. 2018;132: 1225–1240.

10. Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, et al. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. Cancer Cell. 2016;30: 404–417.

11. Shiozawa Y, Malcovati L, Gallì A, Sato-Otsubo A, Kataoka K, Sato Y, et al. Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. Nat Commun. 2018;9: 3649.

12. Bondu S, Alary A-S, Lefèvre C, Houy A, Jung G, Lefebvre T, et al. A variant erythroferrone disrupts iron homeostasis in -mutated myelodysplastic syndrome. Sci Transl Med. 2019;11. doi:10.1126/scitranslmed.aav5467

13. Inoue D, Chew G-L, Liu B, Michel BC, Pangallo J, D'Avino AR, et al. Spliceosomal disruption of the non-canonical BAF complex in cancer. Nature. 2019;574: 432–436.

14. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB. Identification and analysis of alternative splicing events conserved in human and mouse. Proc Natl Acad Sci U S A. 2005;102: 2850–2855.

15. Bernard E, Nannya Y, Hasserjian RP, Devlin SM, Tuechler H, Medina-Martinez JS, et al. Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes. Nat Med. 2020;26: 1549–1556.

16. Papapetrou EP. Modeling myeloid malignancies with patient-derived iPSCs. Exp Hematol. 2019;71: 77–84.

17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

18. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

19. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015;4: 1521.

20. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

21. Reyes A, Anders S, Weatheritt RJ, Gibson TJ, Steinmetz LM, Huber W. Drift and conservation of differential exon usage across tissues in primate species. Proc Natl Acad Sci U S A. 2013;110: 15377–15382.

22. Van den Berge K, Soneson C, Robinson MD, Clement L. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. Genome Biol. 2017;18: 151.

23. Shen S, Park JW, Lu Z-X, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111: E5593–601.

24. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16: 85–97.

25. Agamah FE, Bayjanov JR, Niehues A, Njoku KF, Skelton M, Mazandu GK, et al. Computational approaches for network-based integrative multi-omics analysis. Front Mol Biosci. 2022;9: 967205.

26. Veiga D. maser. Bioconductor; 2018. doi:10.18129/B9.BIOC.MASER

27. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

28. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10: 1213–1218.

29. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44: W160–5.

30. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.

31. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics. 2011. doi:10.1214/11-aoas466

32. Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, et al. Cancer-Associated SF3B1 Hotspot

Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. Cell Rep. 2015;13: 1033–1045.

33. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48: 1193–1203.

34. Wang Y, Xu X, Maglic D, Dill MT, Mojumdar K, Ng PK-S, et al. Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer. Cell Rep. 2018;25: 1304–1317.e5.

35. Harvey KF, Zhang X, Thomas DM. The Hippo pathway and human cancer. Nat Rev Cancer. 2013;13: 246–257.

36. Ma S, Meng Z, Chen R, Guan K-L. The Hippo Pathway: Biology and Pathophysiology. Annu Rev Biochem. 2019;88: 577–604.

37. Wang T, Pine AR, Kotini AG, Yuan H, Zamparo L, Starczynowski DT, et al. Sequential CRISPR gene editing in human iPSCs charts the clonal evolution of myeloid leukemia and identifies early disease targets. Cell Stem Cell. 2021;28: 1074–1089.e7.

38. Wesely J, Kotini AG, Izzo F, Luo H, Yuan H, Sun J, et al. Acute Myeloid Leukemia iPSCs Reveal a Role for RUNX1 in the Maintenance of Human Leukemia Stem Cells. Cell Rep. 2020;31: 107688.

39. Chang C-J, Kotini AG, Olszewska M, Georgomanoli M, Teruya-Feldstein J, Sperber H, et al. Dissecting the Contributions of Cooperating Gene Mutations to Cancer Phenotypes and Drug Responses with Patient-Derived iPSCs. Stem Cell Reports. 2018;10: 1610–1624.

40. Kotini AG, Chang C-J, Boussaad I, Delrow JJ, Dolezal EK, Nagulapally AB, et al. Functional analysis of a chromosomal deletion associated with myelodysplastic syndromes using isogenic human induced pluripotent stem cells. Nat Biotechnol. 2015;33: 646–655.

41. Dolatshad H, Pellagatti A, Liberante FG, Llorian M, Repapi E, Steeples V, et al. Cryptic splicing events in the iron transporter ABCB7 and other key target genes in SF3B1-mutant myelodysplastic syndromes. Leukemia. 2016;30: 2322–2331.

42. Nikpour M, Scharenberg C, Liu A, Conte S, Karimi M, Mortera-Blanco T, et al. The transporter ABCB7 is a mediator of the phenotype of acquired refractory anemia with ring sideroblasts. Leukemia. 2013;27: 889–896.

43. Kanagal-Shamanna R, Montalban-Bravo G, Sasaki K, Darbaniyan F, Jabbour E, Bueso-Ramos C, et al. Only SF3B1 mutation involving K700E independently predicts overall survival in myelodysplastic syndromes. Cancer. 2021;127: 3552–3565.

44. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16: 284–287.

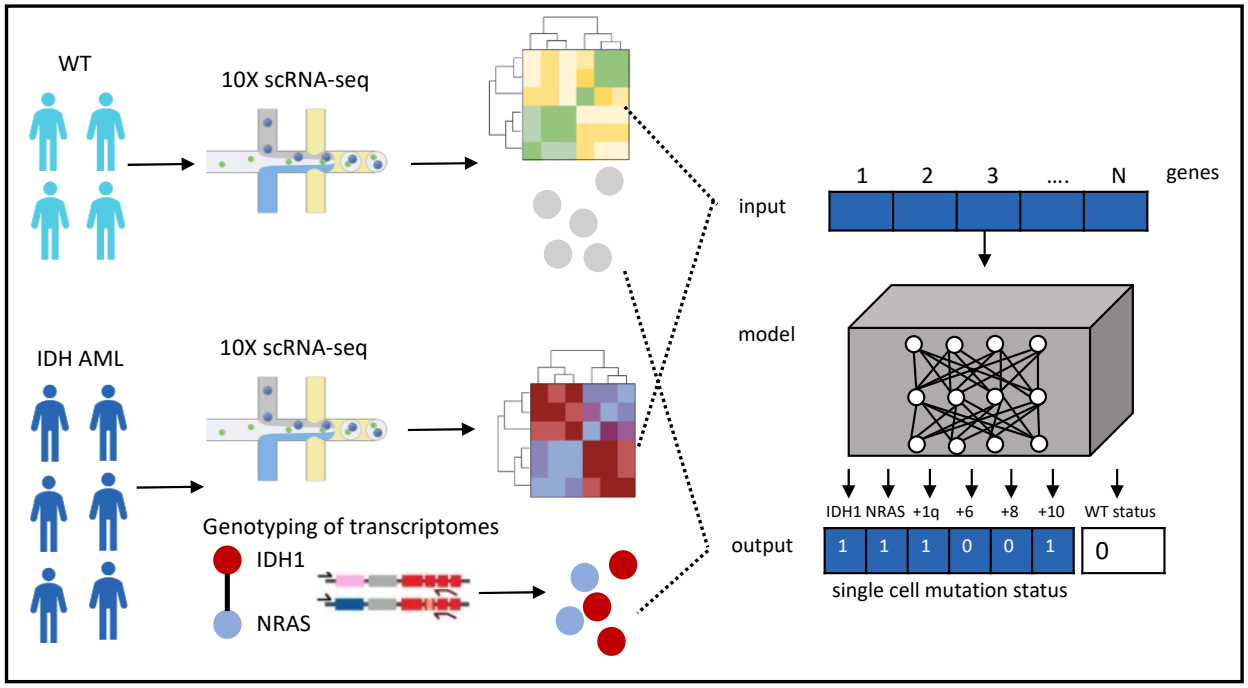
45. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38: 576–589.

46. Teo SH, Tan M, Knight L, Yeo SH, Ng I. Pericentric inversion 9--incidence and clinical significance. Ann Acad Med Singapore. 1995;24: 302–304.

47. Lee S-G, Park TS, Lim G, Lee K-A, Song J, Choi JR. Constitutional pericentric inversion 9 and hematological disorders: a Korean tertiary institution's experience over eight years. Ann Clin Lab Sci. 2010;40: 273–277.

Chapter 3

Predicting single cell genotypes from single cell expression profiles in AML using deep learning



3.1. Chapter abstract

Acute myeloid leukemia (AML) is an aggressive hematologic malignancy composed of a mixture of genotypically, phenotypically and functionally diverse cell populations including wild-type (WT) cells. The generation of high throughput single cell gene expression and mutational profiles in AML enables the deployment of deep learning frameworks for gaining insights on how genotypic changes are associated with disease phenotypes. However, the question of whether the single cell gene expression patterns together with the computational power of neural networks have the capacity to predict a cell's genotype remains unclear. In this study, we train two supervised deep learning models to predict the cell's malignant or wild-type (WT) status as well as the mutational status of specific genomic abnormalities in a binary and multi-class multi-label setting respectively, based on single cell RNA sequencing data from 6 *IDH1/2*-mutated AML patients and 4 healthy individuals. In the independent test sets, the binary classification model achieved an accuracy of 98% while the multi-class multi-label model achieved a macro-average AUC ROC of 0.84. Moreover, applying black box feature selection on the trained networks identified genes involved in biological processes and pathways of reported significance in AML, such as apoptosis and NF-kB related signaling pathways. Overall, this study proposes two deep learning tasks for the prediction of single cell genotypic profiles from single cell expression data and showcases how the trained models can be used for the derivation of biologically related signals.

3.2. Introduction

Acute myeloid leukemias (AML) are aggressive hematologic malignancies with acute onset, rapid progression and poor patient outcomes[1–3]. The pathogenesis of AML is underlied by the serial acquisition of gene mutations in hematopoietic stem and progenitor cells. These mutations impair normal cell regulation and result in a block in the differentiation of myeloid precursors towards more mature myeloid cell types. Thus, AML is characterized by the increased proliferation and accumulation of abnormal immature myeloid progenitor cells (blasts) in the bone marrow (BM) and blood[4,5].

Large genomic population studies have shown that AML is genetically heterogeneous and is defined by the gradual accumulation of multiple gene mutations with specific patterns with regards to mutation order and co-occurrence[6,7]. This genetic and clonal diversity in AML imposes one of the biggest challenges in treating and ultimately curing the disease. Understanding how specific gene mutations result in distinct populations of cells and elucidating how clones with distinct differentiation characteristics lead to malignancy, cannot be achieved by bulk sequencing approaches. The latter methods capture tissue related information and cannot provide information at the cellular level, thus limiting our understanding on how mutations drive AML pathogenesis and how genotypic changes are reflected in clone specific gene expression signals. In contrast to bulk approaches, single cell sequencing technologies provide insights towards the characterization of intra-patient cell diversity and clonal heterogeneity as well as empower the study of gene expression profiles between individuals with different conditions at a single cell resolution. The wealth of data generated by single cell sequencing has also created the opportunity for the design and implementation of deep learning

methods that explore the high-order structure of the data, embed cells on lower dimensional spaces, identify cell clusters, integrate different modalities and reveal different aspects of biological signals[8–10].

In AML, single cell gene expression and single cell genotypes can be combined to assign cells into distinct clones across the AML phylogeny. This allows the identification of different cell populations, the analysis of the interactions between cells and the establishment of relationships between gene expression heterogeneity and genotypically distinct subclones. In particular, Petti et al.[11] evaluated the capacity of using scRNA-sequencing reads from cryopreserved BM cells from AML patients to detect, at a single cell level, a set of somatic variants called by enhanced whole-genome sequencing (eWGS) on the same samples. Exploiting the detected single nucleotide variants (SNVs) of each cell, Petti et al. distinguished tumor and normal cells and examined the cell composition of identified gene expression clusters at the phenotypic and mutational level in an unsupervised setting. In another study, Van Galen et al.[12] utilized nanowell-based technology to collect single cell gene expression and mutational profiles from AML patients as well as healthy individuals and deployed random forests to predict the WT or malignant status of each cell in a two-step approach. First, Van Galen et al. applied a random forest to assign mutated AML cells a label based on a defined set of WT cell types and then separated malignant and WT cells by training a second random forest model to classify cells to WT or malignant cell type labels.

Inspired by the work of Petti et al. and Van Galen et al. on AML as well as by the applicability of deep learning to single cell data, this paper first, introduces a deep learning framework to classify AML malignant and WT cells in a binary setting and second, attempts to predict the single cell mutational status of specific genes in a multi-label supervised setting. Both tasks are performed using 1. single cell gene expression profiles from scRNA-seq data from 6 *IDH1/2*-mutated AML patients and 4 WT individuals and, 2. genotype labels produced experimentally from applying the method of Genotyping of Transcriptomes (GoT)[13] on the same samples. Through this deep learning approach, we integrate single cell gene expression with single cell genotypes in AML and enable the identification of genes that play a significant role in these classification tasks. Our main contributions in this study are: i) deploying a feedforward neural network to classify cells to WT or malignant in a supervised approach using as input single cell expression profiles from diagnostic AML patient samples and healthy individuals, ii) deploying a feedforward neural network to predict the single cell mutational status of a specific set of genes and chromosomal abnormalities for a single patient in a supervised multi-class, multi-label setting using the same single cell gene expression profiles, and iii) identifying features-genes that are important for both classification outcomes by applying the holdout randomization test (HRT)[14] on the trained models.

3.3. Data

Study cohort

The study cohort consists of 6 AML patients and 4 healthy individuals (N-01, N-02, N-03, N-04) from the paper of Sirenko et al (in review)[15]. These 6 AML patients harbored clonal *IDH1* (patient IDH1i-01, patient IDH1i-

02, patient IDH1i-03) or *IDH2* (patient IDH2i-01, patient IDH2i-02, patient IDH2i-03) mutations as well as co-mutations in one or more of the most commonly co-mutated genes in *IDH1/2* AML; *NPM1*, *NRAS*, *KRAS*, *SRSF2*, *DNMT3A* (Figure 3.1A). For this cohort, scRNA-seq data were generated from BM and peripheral blood (PB) samples. Next to the single cell gene expression profiles, single cell genotypic information was also available for the 6 AML patients for specific hotspot mutations (*IDH1 p.R132*, *IDH2 p.R172*, *IDH2 p.R140*, *SRSF2 p.P95*, *DNMT3A p.R882*, *NPM1 p.W288*, *NRAS p.G12*, *KRAS p.G12*), as derived from the application of the GoT method on the same samples[13] (Supplemental Table 3.1). Additionally, a set of chromosomal abnormalities (gains in chromosomes 1q [+1q/dupli_chr1], 6 [+6/dupli_chr6], 8 [+8/dupli_chr8], 10 [+10/dupli_chr10] and 14 [+14/dupli_chr14]), as reported from inferCNV at the single cell level, is included[16]. We note that all patient samples were treatment naive and that the generation of the data is not part of the current thesis.

Data preprocessing

The raw scRNA-seq fastq files were aligned to the GRCh37 assembly and single cell gene expression counts were generated using CellRanger v3.1[17]. We enhanced the quality of the AML and healthy datasets by performing a series of quality control steps with scanpy[18] including the removal of cells with less than 200 expressed genes, of genes expressed in less than 3 cells and of cells with mitochondrial content more than 20%. Lastly, based on cell type annotations, we remove any lymphoid-related cells. To bring the single cell gene expression values of the AML and healthy datasets to the same scale, we applied the sctransform normalization[19] using Seurat[20]. Then, we integrated the healthy and AML single cell gene expression profiles into a unified dataset, first by identifying correspondences between pairs of single cells from the two datasets, and second by transforming the gene expression values of the datasets into a common space[20]. This unified dataset contains in total 61,091 myeloid cells (Figure 3.1B). Applying UMAP on the top 30 principal components and coloring these cells on the 2D UMAP space based on the dataset of origin, we can visually identify both overlapping and non-overlapping single cell profiles between WT and AML (Figure 3.1C). Thus, we can rationally ask if it is possible, through deep learning, to predict 1) each cell's status (malignant or WT) and 2) each cell's genomic abnormalities based on single cell gene expression values. To do that, any cell, from the AML patients, with at least one hotspot mutation or chromosomal abnormality was labeled as malignant while every cell from the WT individuals was labeled as WT (Figure 3.1D). Therefore, the total number of cells used for the downstream training, validation and testing reached a total of 50,026, 35,314 of which were malignant and 14,712 of which were WT (Figure 3.1E, Supplemental Table 3.2).

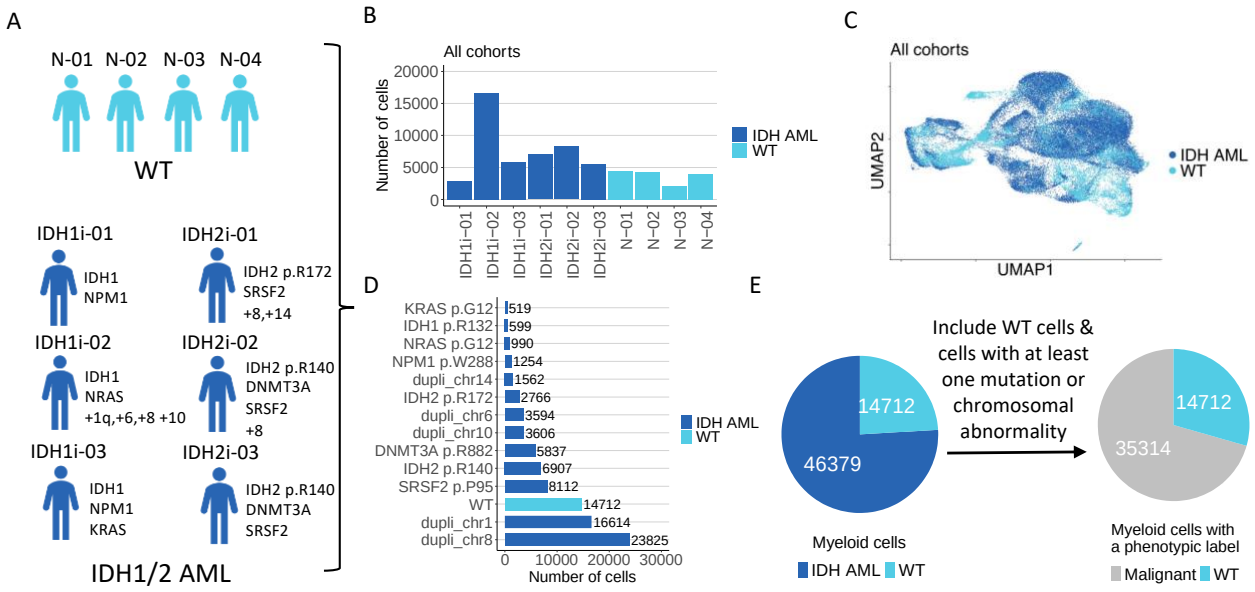


Figure 3.1. Study cohort and data characteristics. A) Study cohort composed of 6 AML patients and 4 healthy individuals from Sirenko et al (in review). The chromosomal abnormalities and genotyped mutated genes used in the study are listed next to each patient. B) Barplot showing the number of cells with gene expression data per individual, colored by dataset. C) UMAP of all cells based on their expression profiles, colored by dataset. D) Horizontal barplot showing the number of cells with genotypic labels per genomic abnormality, colored by dataset. E) Pieplots showing the total number of cells per dataset (left) and the total number of cells used for the training, validation and testing of the models.

3.4. Methods

In this study we develop a binary classification and a multi-class, multi-label model to predict the cell's malignant or WT status and the mutational status of specific genomic abnormalities correspondingly, using the single cell gene expression profiles of the integrated AML and healthy datasets. Let $X \in R^{M \times N}$ be the single cell gene expression matrix, where M = 50,026 is the total number of malignant and WT cells, while $N = 1{,}000$ is the number of genes with the highest expression variance across cells (Figure 3.2A). Additionally, let G be the $M \times L$ matrix of genotype labels, where $L = 13$ is the total set of hotspot mutations from GoT (n=8) and chromosomal abnormalities from inferCNV (n=5). A value of $G_{i,j} = 1$ denotes that the genomic abnormality j is present in cell i, a value of $G_{i,j} = 0$ denotes that the genomic abnormality j is not present in cell i, while NA denotes that there is not enough information for assigning a 0 or 1 (e.g. dropout in GoT) for the genomic abnormality j in cell i (Figure 3.2A). Since $M = 50{,}026$ is the total number of malignant and WT cells, there is no single row in G that contains only NAs. Lastly, the binary vector S of size M indicates if each cell i is malignant ($S_i = 1$) or WT ($S_i = 0$, Figure 3.2A).

Binary classification model

First, we deploy a feedforward neural network of 3 hidden dense layers (sizes 512, 512, 64 respectively) to predict if a cell is WT or malignant (S_i) based on its expression profile X_i (Figure 3.2B). In particular, the

network receives as input each single cell gene expression profile, creates a latent representation of this profile through a series of hidden non-linear operations and eventually outputs a probability estimate that indicates how likely each cell is to be malignant. The output $H_i^{(l)}$ of each hidden layer $l \in \{1,2,3\}$ for cell i is defined as:

$$H_i^{(l)} = ReLU\left(H_i^{(l-1)} \times W^{(l)} + b^{(l)}\right), i \in \{1..M\} \qquad (1)$$

where $W^{(l)}$ and $b^{(l)}$ are the transposed learnable weights and the bias of layer l, respectively, and $H^{(0)} = X$. The output O_i of the model for cell i is defined as:

$$O_i = Sigmoid(H_i^{(3)} \times W^{(out)} + b^{(out)}) \qquad (2)$$

where $W^{(out)}$ and $b^{(out)}$ are the transposed learnable weights and the bias of the output layer, respectively. The proposed end to end model is trained in a supervised fashion by minimizing the binary cross-entropy loss L^{BCE}:

$$L_i^{BCE} = -[S_i \cdot logO_i + (1 - S_i) \cdot log(1 - O_i)], i \in \{1..M\} \qquad (3)$$

where L_i^{BCE} is the binary cross entropy loss for cell i and O_i is the output for cell i. Furthermore, to reduce overfitting and improve generalization, during training we used dropout regularization after the output $H^{(l)}$ of each hidden layer. Upon the completion of training and the evaluation of the model performance on a holdout unseen test set, we identify which genes are important for the classification decision by treating the trained model as a black box and applying the HRT method[14] for feature selection (Figure 3.2B). HRT handles the feature selection task as a hypothesis testing problem and selects the features that are relevant to the outcome by performing conditional independence tests. Under the null hypothesis (feature conditionally independent of the outcome given all other features), the feature is irrelevant to the outcome. When the null hypothesis is rejected, then the respective feature is accounted as a discovery[14].

Multi-class multi-label classification model

Next, we extend the previous binary classification model aiming to also predict the state (mutated or not mutated) of each hotspot position and chromosomal aberration. For this task, we develop a patient-specific multi-class, multi-label model for the patient with the largest amount of data (IDH1i-02) (Figure 3.1B, Supplemental Table 3.3). This model retains the feedforward architecture deployed earlier, but produces 7 different outputs (Figure 3.2C), 6 of which correspond to the presence of a hotspot mutation or a chromosomal abnormality ($j \in A = \{IDH1, NRAS, dupli\_chr1, dupli\_chr6, dupli\_chr8, dupli\_ch10\}$) and the last one ($j \in \{WT_{status}\}$) corresponds to the genotypic status of the cell (WT or malignant). The data used for training, validating and testing this model consists of all malignant cells from patient IDH1i-02 (n=16,614) as well as all WT cells (n=14,712) of the cohort. We denote the sum of these cells as K ($K = 31{,}326$). Similarly to the binary classification model, this extended network receives as input each single cell gene expression profile X_i and creates a latent representation of this profile through a series of hidden non-linear operations. Eventually, the network produces a probability estimate for each output $j \in A$ that

indicates how likely each mutation or chromosomal abnormality is to be present in each cell i, as well as $j \in \{WT_{status}\}$ that indicates how likely each cell i is to be WT (contrary to the binary classification model in which the output was estimating the probability of the cell being malignant). The output $H_i^{(l)}$ of each hidden layer $l \in \{1,2,3\}$ (sizes 512, 512, 512 respectively) for cell i is as in equation (1) while the model output can be written as:

$$O_{i,j} = Sigmoid(H_i^{(3)} \times W_j + b_j), i \in \{1..K\}, j \in A \cup \{WT_{status}\} \qquad (4)$$

where $O_{i,j}$ is the probability estimate for output j for cell i, and W_j and b_j are the transposed weight parameters and the bias of the output layer, respectively, that correspond to each output j. The proposed multi-class, multi-label end to end model is trained in a supervised fashion by minimizing the sum (L^{TOT}) of two binary cross-entropy losses; L^{BCE} that penalizes any deviation of the model output from the true output labels and L^{OVL} that penalizes cells with high output probabilities of having any mutation and being predicted as WT at the same time or vice versa (low probabilities of any mutation and being predicted as malignant). These losses are defined as:

$$L_{i,j}^{BCE} = -w_j \left[T_{i,j} \cdot log O_{i,j} + \left(1 - T_{i,j}\right) \cdot log(1 - O_{i,j}) \right], j \in A \cup \{WT_{status}\}, i \in \{1..K\} \qquad (5)$$

$$L_i^{OVL} = -[max\,(O_{i,j \in A}) \cdot log(1 - O_{i,WT_{status}}) + (1 - max(O_{i,j \in A})) \cdot log(O_{i,WT_{status}})], i \in \{1..K\} \qquad (6)$$

where $O_{i,j}$ is the output prediction j for cell i,

$$T_{i,j} = \begin{cases} 1 - S_i, & if\ j \in \{\mathrm{WT_{status}}\} \\ G_{i,j}, & if\ j \in A \end{cases} \qquad (7)$$

and $w_j \in R$ is the weight of each label {0,1,NA} for output j defined as:

$$w_j = \begin{cases} 0, & if\ T_{i,j} = NA \\ v_0 \in R, if\ T_{i,j} = 0 \\ v_1 \in R, if\ T_{i,j} = 1 \end{cases} \qquad (8)$$

If the true label for output j is NA, the model output j cannot be evaluated for its correctness and thus based on equation (8), there isn't any loss component from output j contributing to L^{BCE}. The nonzero w_j (v_0 for label 0 and v_1 for label 1) are computed based on the prevalence of the labels {0,1} in G_j. The total loss for cell i is defined as:

$$L_i^{TOT} = L_i^{OVL} + \frac{\sum_{j \in A \cup \{WT_{status}\}} L_{i,j}^{BCE}}{7}, i \in \{1..K\} \qquad (9)$$

Similarly to the binary classification model, we used dropout regularization after the output $H^{(l)}$ of each hidden layer. Additionally, we apply the HRT method[14] on the trained model to identify the genes that play a significant role in the positive prediction of the NRAS mutation.

3.5. Results

Binary classification model accurately predicts malignant cells from WT cells

For the binary classification task we split the data into training, validation and test sets. The model was trained and optimized based on the training and validation sets and tested for its performance on the holdout test data. Particularly, it was trained for 283 epochs in batches of 64 cells using the Adagrad optimizer with a learning rate of 0.013 and a weight decay of 0.01 (Supplemental Figure 3.1).

Outputs with a probability higher than 0.5 were regarded as positive and represented malignant predictions, whereas the rest were regarded as negative and represented WT predictions. The evaluation of the model performance on the test set (Figure 3.3A) showed that the model has the capacity to separate malignant from WT cells with an accuracy of 98%, precision of 98% and recall of 99%, resulting in 70.19% of the test set cells being True Positives (TP, malignant cells that were correctly classified as such), 28.28% being True Negatives (TN, WT cells that were correctly classified as such), 1.13% being False Positives (FP, WT cells that were classified as malignant) and 0.4% being False Negatives (FN, malignant cells that were classified as WT).

Additionally, running the HRT method twice on the test set in the context of identifying the most important features for this classification task, led to the selection of 58 common genes between the two runs with a Benjamini-Hochberg (BH) adjusted p-value<0.05. To find out if these genes share similar biological functions or participate in the same biological processes, we performed a gene ontology and pathway enrichment analysis[21] on this selected set of genes which showed the enrichment of processes related to apoptosis (BH adjusted p-value = 0.009, e.g. *MCL1*, *HMGB2*) as well as of the TGF-beta signaling pathway (BH adjusted p-value = 0.005, e.g. *ID1*, *JUNB*) (Figure 3.3B). We further used this trained model to search for the presence of WT-like cells within the AML patients (Figure 3.3C). Applying the model to the cells of the AML patients that were not part of the training, validation and test sets, we find a small portion of cells within each patient that present a phenotype similar to that of the WT cells (WT-like). In total, these WT-like predictions are the 4.1% of this cell-set and 56% of them correspond to myeloid differentiated cells (Figure 3.3D), which may have escaped the differentiation block and reached myeloid maturation. Notably, only 8% of the malignant predicted cells of this set correspond to myeloid differentiated ones, indicative of the differentiation block that characterizes the disease.

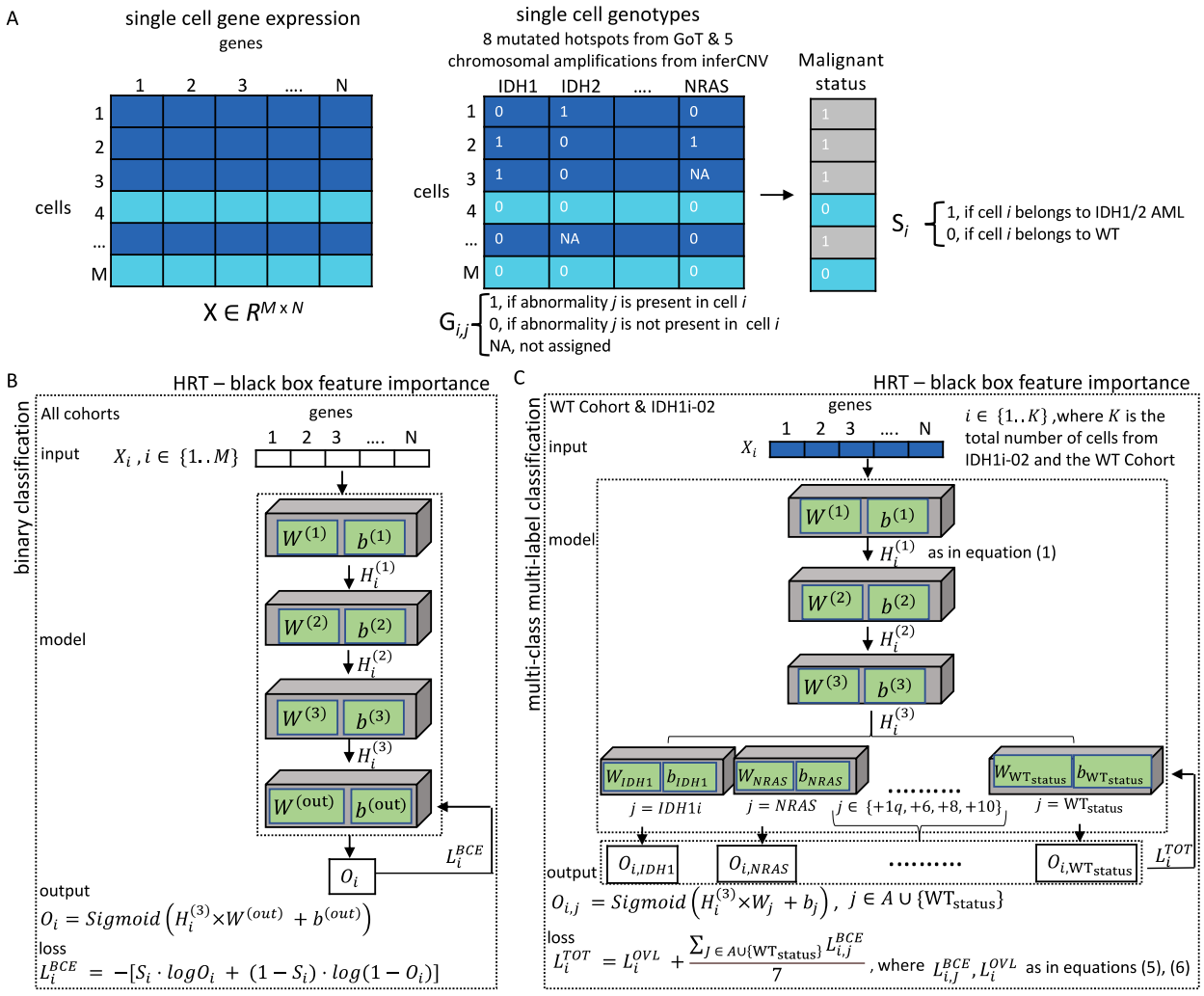


Figure 3.2. Data representation, frameworks and use of model equations. A) 2D single cell gene expression matrix $X \in R^{M\times N}$ (left), 2D single cell genotype matrix G, $G_{i,j} \in \{0,1,NA\}, i \in \{1..M\}, j \in A \cup \{WT_{status}\}$ (middle) and 1D matrix S, $S_i \in \{0,1\}$, $i \in \{1..M\}$ representing if a cell is malignant or WT (right). B) Binary classification model and use of model equations in a forward pass example. The model receives as input single cell expression profiles X_i, produces a series of latent representations $H_i^{(l)}, l \in \{1,2,3\}$ and outputs the probability ($O_i, i \in \{1..M\}$) of a cell i being malignant using the sigmoid activation function. The binary cross entropy loss (L_i^{BCE}) between the output probability O_i and the true label S_i is used during training. $W^{(l)}$ are the transposed weight parameters and $b^{(l)}$ is the bias of layer $l, l \in \{1,2,3\}$, while $W^{(out)}$ and $b^{(out)}$ are the transposed weight parameters and the bias of the output layer, respectively. C) Multi-class multi-label model and use of model equations in a forward pass example. The model receives as input single cell expression profiles X_i, produces a series of latent representations $H_i^{(l)}, l \in \{1,2,3\}$ and outputs the probabilities $O_{i,j}$ of a cell $i \in \{1..K\}$ (K is the total number of cells from IDH1i-02 and the WT Cohort) harboring each genomic abnormality $j \in A$ and the probability of cell i being WT ($j \in \{WT_{status}\}$). The model is developed for patient IDH1i-02 using this patient's cells and the WT cells from the healthy individuals too. Besides the binary cross entropy loss $L_{i,j}^{BCE}$ between $O_{i,j}$ and the true label for $j \in A \cup \{WT_{status}\}$, the training of this model also uses L_i^{OVL}, a binary cross entropy loss between the maximum probability of a genomic abnormality being present and the probability of the cell i being malignant. The total loss L_i^{TOT} for cell i during a training iteration is the sum of L_i^{OVL} and the average of $L_{i,j}^{BCE}$

across all outputs j. $W^{(l)}$ and $b^{(l)}$ are the transposed weight parameter and the bias of layer l , $l \in \{1,2,3\}$ respectively, while W_j and b_j are the transposed weight parameters and the bias of the output layer, respectively, that correspond to each output j. The HRT method is applied on both trained models for black box feature selection. M: total number of cells; N: number of input features (genes); A is the set of genomic abnormalities.

Multi-label classification model effectively predicts NRAS mutational status

For the multi-class multi-label classification framework, we split the data to training, validation and test, ensuring at the same time that all sets contain cells with genomic abnormalities for every output $j \in A$ (there is at least one $i \in \{1..K\}$ for which $G_{i,j}$= 1, $j \in A$). The model was trained and optimized based on the training and validation sets and tested for its performance on the unseen test data similarly to the binary classification model. In this case, the model was trained for 68 epochs in batches of 128 cells using the SGD optimizer with a learning rate of 0.086 and a weight decay of 0.001 (Supplemental Figure 3.2).

In this multi-class multi-label setting we do not use the 0.5 probability threshold to determine positive and negative outcomes, but we adjust the classification threshold for each output separately. The threshold selected for each output was the one achieving the highest f1 score on the precision-recall curve of the validation set. Therefore, each output j with probability estimate higher than the output-specific threshold was regarded as positive (predicted presence of genomic abnormality $j \in A$, predicted WT for $j \in \{WT_{status}\}$) while every output with probability estimate below that threshold was regarded as negative (predicted absence of genomic abnormality for $j \in A$, predicted malignant for $j \in \{WT_{status}\}$).

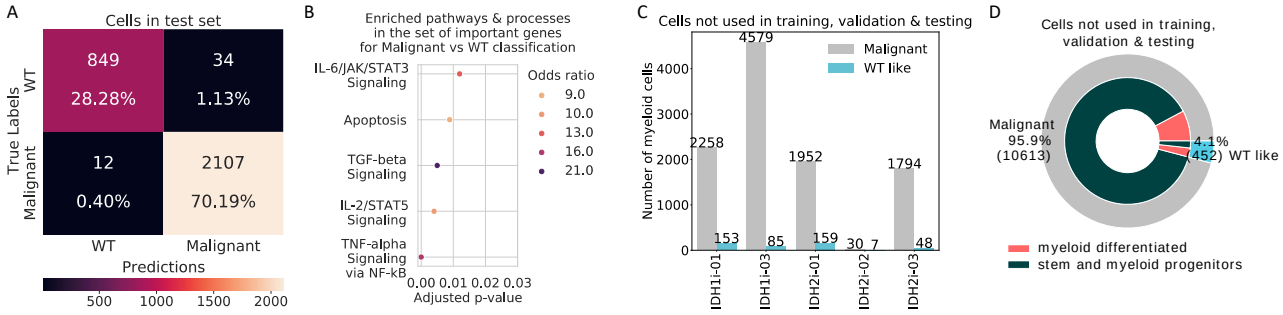


Figure 3.3. Results of the binary classification model. A) Confusion matrix of the binary classification model on the test set. B) Dotplot showing AML related biological processes and pathways that are enriched in the set of selected genes resulting from the application of HRT for the test set on the binary classification model. The y axis presents the terms, the x axis shows the statistical significance for each term and the color of the dots indicates the odds ratio of the terms. C) Barplot showing the predictions of the binary classification model on the set of cells not used in the training, validation and testing of the model (cells without available genotypic information). The x-axis shows the patients and the color indicates the model prediction. D) Nested donut plot showing the model predictions on the cells not used in the training, validation and testing (outer) along with the indication of the myeloid maturation stage (inner).

Similar to the binary classification case, this model also achieves the (98% of the predictions are correct) separation between malignant and WT cells on the test set (Figure 3.4A). Complementary to this, we also note for clarity that the model does also predict absence of genomic abnormalities ($j \in A$) in almost all (in

2262 out of the 2272) of the WT predicted cells. As far as genomic abnormalities ($j \in A$) are concerned, the model performance is assessed only on the malignant cells of the test set. In this context, the model achieved near optimal results on the prediction of the chromosomal abnormalities, especially of dupli_chr6 and dupli_chr10 (Figure 3.4A), achieving an AUC ROC higher than 0.96 for all (Figure 3.4B, for dupli_chr1, no AUC ROC is computed as all cells had a chromosome 1 gain). Regarding the mutations, the model demonstrated a considerable performance for subclonal *NRAS* achieving an AUC ROC of 0.83 (Figure 3.4B, Supplemental Table 3.4), in contrast to its limited capacity to correctly predict the mutational status of clonal *IDH1*. The latter may be due to the lack of *IDH1* WT cells from the cohort and consequently the training set (Figure 3.4C, Supplemental Table 3.3).

Given the performance of the model on *NRAS* as well as the fact that *NRAS* is one of the most commonly co-mutated genes in AML[7], we applied the HRT method twice on the cells of the test set that were predicted to have an *NRAS* mutation. This led to the selection of 82 common genes between the two runs (BH adjusted p-value<0.05) as important for the positive *NRAS* prediction. Gene ontology and pathway enrichment analysis on this set of genes showed, among others, association of these genes with inflammatory responses (BH adjusted p-value = 0.006, e.g. *RNF144B, TLR2*) as well as the TNF-alpha signaling via NF-kB (BH adjusted p-value = 0.00003, e.g. *BTG2, SAT1*) and IL-2/STAT5 (BH adjusted p-value = 0.028, e.g. *XBP1, HOPX*) signaling pathway (Figure 3.4D). We note that the outcome of the enrichment analysis shows which terms are over-represented in the set of genes derived from HRT and does not imply the up or down-regulation of the processes/pathways and their matched genes.

3.6. Discussion

In this study, in the context of *IDH1/2* AML and using as input single cell gene expression profiles, we propose a binary classification and a multi-label deep learning model to predict the Malignant or WT status of single cells and their specific genomic abnormalities, respectively. To develop these models, we integrated the single cell gene expression and genotypic data from 6 *IDH1/2*-mutated AML patients and 4 healthy individuals from Sirenko et al (in review). Motivated by the work of Van Galen et al. on AML, who deployed cell type labeling and two random forest models to separate between malignant and WT cells, this paper leverages the computational power of deep learning models to not only predict malignant vs WT cell status, but also identify the mutational status of genomic abnormalities for a single patient despite missing data. In particular, the multi-class multi-label architecture aims to shape internal cell representations through the sharing of information learned from different outputs, recovering that way for the excessive absence of mutation labels for some genomic abnormalities.

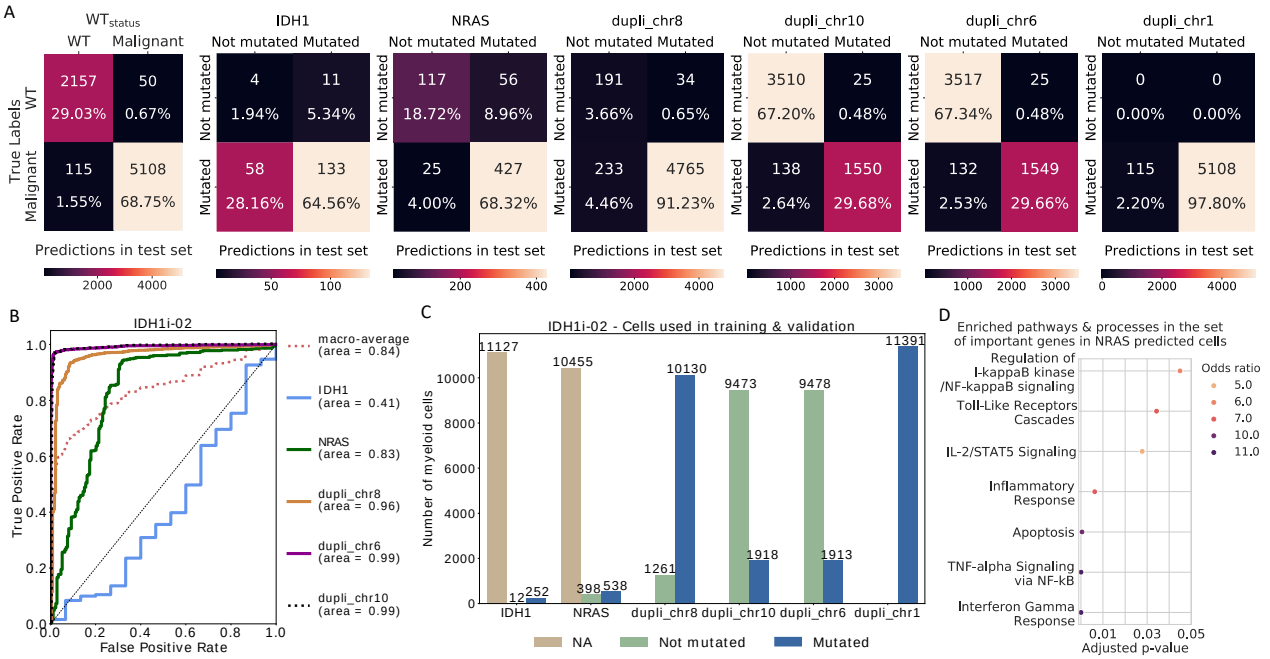


Figure 3.4. Results of the multi-label classification. A) Confusion matrix for each output of the multi-class multi-label model on all test set cells for $j \in \{WT_{status}\}$ and on the test set cells of patient IDH1i-02 for $j \in A$ in which $G_{i,j} \neq NA$. B) ROC curve for each output $j \in A$ of the multi-class multi-label model along with the respective AUC on the test set. C) Barplot showing the number and mutational status of the cells from patient IDH1i-02 used for the training and validation of the multi-class multi-label model. The x axis shows the genomic abnormalities and the color indicates the presence, absence or unavailability of the label. D) Dotplot showing the biological processes and pathways that are enriched in the set of selected genes resulting from the application of HRT on the multi-class, multi-label model for the test set cells with a predicted *NRAS* mutation. The y axis presents the terms, the x axis shows the statistical significance for each term and the color of the dots indicates the odds ratio of the terms.

Both deep learning models showed similarly excellent performance in classifying malignant from WT cells (98 % correct test set predictions from both models) while the multi-class multi-label model for patient IDH1i-02 presented remarkable results in predicting the status of chromosomal abnormalities and of *NRAS*, in contrast to *IDH1*. We believe that the superior performance of the model on *NRAS* compared to *IDH1* is related to the limited genotyping efficiency of *IDH1* in the Sirenko et al. data, but it could also depend on the different acquisition stages of the two mutations. Specifically, *IDH1* is a clonal mutation, while the acquisition of *NRAS* in patient IDH1i-02 is a later event that might have such an effect on the gene expression profiles of the mutated cells that makes them more easily distinguishable for the model. The fact that *NRAS* is a subclonal event with higher genotyping efficiency than *IDH1*, provided for patient IDH1i-02 a quite balanced set of NRAS genotyped cells in which 37% is *NRAS* WT (Figure 3.4C). That allowed us to train a patient-specific model with considerable results for *NRAS*. Excluding the chromosomal abnormalities, we note that within each of the other patients of the cohort, there were at least two mutations in which the gene specific status was WT for less than 30% of the cells.

The superior performance on *NRAS* demonstrates that this classification task is optimal when addressing cells with a spread of mutant and WT representations such as sub-clones. This is of significant translational and clinical relevance as it is often such emerging subclones that carry mutations that confer resistance to treatment, and that seed disease relapse and progression. This will allow one to characterize the biological determinants of specific gene mutations across cell lineages that result in treatment resistance and disease progression as opposed to phenotyping disease initiating mutations that are present in all the cells (e.g. *IDH1/2*).

In the context of obtaining a better understanding of the classification decisions and deep learning model behavior, white box interpretability methods have recently gained much attention[22,23]. However, a number of those (e.g. Integrated Gradients, DeepLift, SHAP variants, Feature Ablation and Occlusion) are based on input baselines[23], the setting of which in the context of the heterogeneous single cell gene expression profiles is tricky and might influence the resulting interpretation. Thus, by applying the HRT feature selection method, we show how these trained models can be treated as black boxes for recovering biologically related associations between the model outputs and the input genes. The HRT results in the binary classification model show that the selected features are, among others, enriched in the TGF-beta signaling pathway, the significance of which in the context of AML has been previously reported[24,25]. The HRT results for *NRAS* predictions in the multi-class, multi-label setting present an association with inflammation related processes, the role of which in AML and other hematologic malignancies has also been previously examined[26,27]. We note that given the shared input between the two models as well as the fact that the positive *NRAS* predictions are also malignant ones, there is an overlap between the associated biological terms in both models. Apoptotic processes and NF-kB related signaling pathways, which have central impact on the cellular functions in AML[28,29], were, among others, associated with the selected genes of both classification tasks.

To conclude, this study first deployed two deep learning frameworks that integrate single cell gene expression profiles and genotypes, and second leveraged them to extract biological signals related to disease. To develop these models we carefully selected a sample set that includes healthy individuals (n=4), as well as patients (n=6) with representative genotypes from *IDH1/2*-mutated AML, for which both scRNA-seq data coupled to single cell genotyping of transcriptomes (GoT) data were generated. While acknowledging the small cohort size of the study, we note that GoT is a very laborious technique that cannot be readily scaled, and the significant cost of the combined assays (scRNA-seq and single cell genotyping) prohibits at present the generation of such data at scale. As part of the analysis, we composed a set of thousands of cells with different genotypes, normalized this set to reduce patient-specific effects, and used it to train frameworks that operate at a single cell level and predict specific genotypes. Future efforts encompassing larger sample sets may focus on enriching the genotype-phenotype associations derived hereby. Additionally, the approach of this study can be extended not only to different scientific questions within the same dataset (e.g. classifying *IDH1* vs *IDH2* cells), but could also be acquired in similar data settings in other diseases. Lastly, future directions could concentrate on overcoming integration issues across single cell RNA-seq data from different sources and technology protocols so that the proposed frameworks can be readily applicable without retraining or calibration on unseen patients on a wide research or clinical scale.

3.7. Supplementary

3.7.1. Supplemental methods

Model training

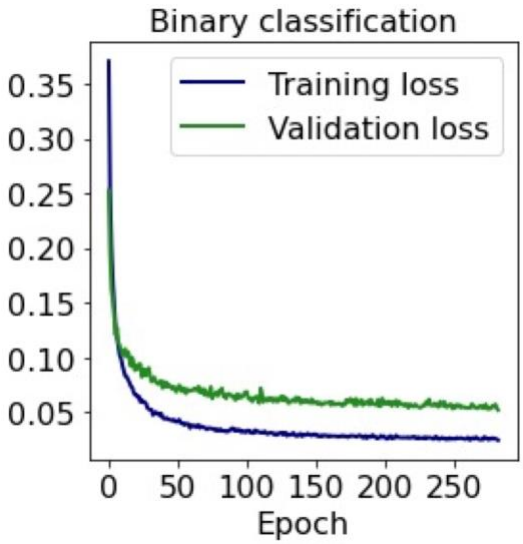
The models were developed in pytorch[30] under python 3.7.1. Hyperparameter tuning (learning rate, weight decay, optimization algorithm) was performed using the Asynchronous Successive Halving Algorithm (ASHA)[31]. Gene enrichment analysis was conducted using GSEApy[32].

Data and code availability

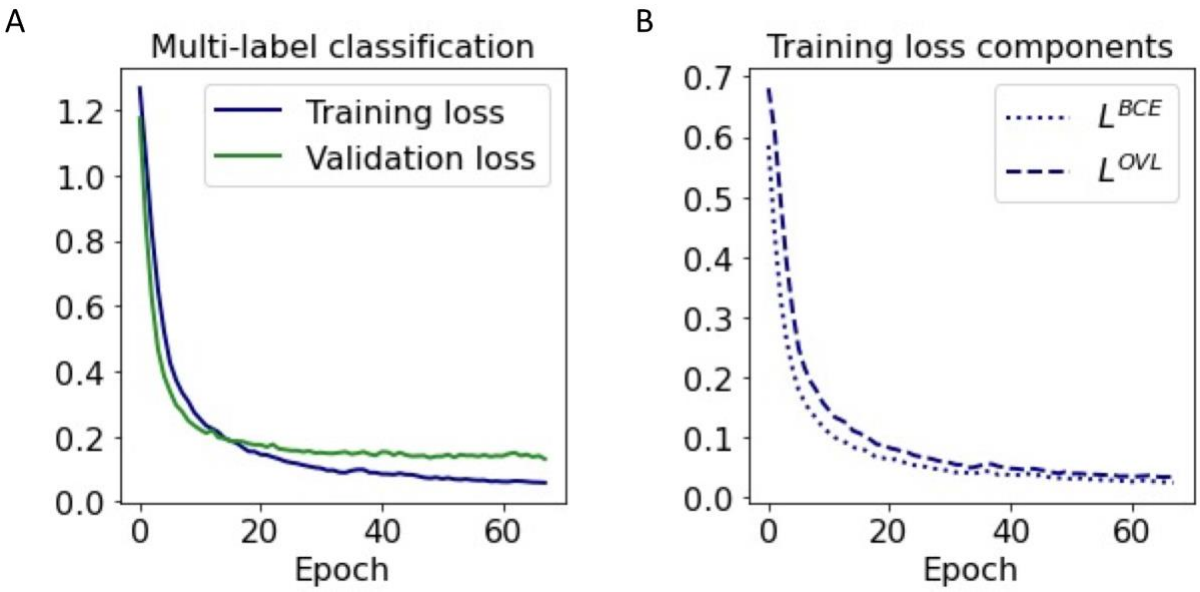
The work presented in this chapter has been published here (Association for Computing Machinery, International Conference on Bioscience, Biochemistry and Bioinformatics 2023 Conference proceedings).

A github repository containing the models' source code is available at the Papaemmanuil lab github page (https://github.com/papaemmelab/Asimomitis_ACM_2023). The single cell dataframes used for the training, validation and testing of the models will become available upon publication of the Sirenko et al.

3.7.2. Supplemental figures



Supplemental Figure 3.1. Curves for the training and validation loss for the binary classification model across epochs.



Supplemental Figure 3.2. A) Curves for the training and validation loss for the multi-label classification model across epochs. B) The two components (L^{OVL} and the average of L^{BCE}) of the training loss for the multi-label classification model.

3.7.3. Supplemental tables

Supplemental Table 3.1. Patient-specific genotypic profiles (genotyped mutations and chromosomal gains) used in the study.

Patient	Genotyped mutations	Gains in chromosomes
IDH1i-01	*IDH1, NPM1*	-
IDH1i-02	*IDH1, NRAS*	1q, 6, 8, 10
IDH1i-03	*IDH1, NPM1, KRAS*	-
IDH2i-01	*IDH2 p.R172, SRSF2*	8,14
IDH2i-02	*IDH2 p.R140, DNMT3A, SRSF2*	8
IDH2i-03	*IDH2 p.R140, DNMT3A, SRSF2*	-

Supplemental Table 3.2. Breakdown, across cohort individuals, of the 50,026 cells used for the tuning and testing of the binary classification model.

Patient	Model tuning	Hold out test set
IDH1i-01	506	45
IDH1i-02	15647	967
IDH1i-03	1181	68
IDH2i-01	4618	311
IDH2i-02	7755	491
IDH2i-03	3488	237
N-01	4141	265
N-02	4044	240
N-03	1911	137
N-04	3733	241

Supplemental Table 3.3. Number of cells per mutational status for each genomic abnormality of patient IDH1i-02, used for the model tuning (training, validation) and the hold out test set of the multi-label framework.

	Model tuning			Hold out test set		
	NA	Not mutated	Mutated	NA	Not mutated	Mutated
IDH1_R132	11127	12	252	5017	15	191
NRAS_G12	10455	398	538	4598	173	452
dupli_chr8	0	1261	10130	0	225	4998
duplic_chr10	0	9473	1918	0	3535	1688
dupli_chr6	0	9478	1913	0	3542	1681
dupli_chr1	0	0	11391	0	0	5223

Supplemental Table 3.4. Performance metrics of the multilabel classification model on the hold-out test set of patient IDH1-02. (TNR: True Negative Rate, ROC AUC: Area Under the ROC curve).

	Accuracy	Precision	Recall	F1 score	TNR	AUC ROC
IDH1_R132	67%	92%	70%	79%	27%	41%
NRAS_G12	87%	88%	94%	91%	68%	83%
dupli_chr8	95%	99%	95%	97%	85%	96%
dupli_chr10	97%	98%	92%	95%	99%	99%
dupli_chr6	97%	98%	92%	95%	99%	99%

3.8. References

1. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016;127: 2391–2405.

2. Burnett A, Wetzler M, Löwenberg B. Therapeutic advances in acute myeloid leukemia. J Clin Oncol. 2011;29: 487–494.

3. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017;129: 424–447.

4. Madan V, Koeffler HP. Differentiation therapy of myeloid leukemia: four decades of development. Haematologica. 2021;106: 26–38.

5. Olsson I, Bergh G, Ehinger M, Gullberg U. Cell differentiation in acute myeloid leukemia. Eur J Haematol. 1996;57: 1–16.

6. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008;456: 66–72.

7. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. N Engl J Med. 2016;374: 2209–2221.

8. Amodio M, van Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods. 2019;16: 1139–1145.

9. Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 2021;12: 1882.

10. Wen H, Ding J, Jin W, Wang Y, Xie Y, Tang J. Graph neural networks for multimodal single-cell data integration. 2022. doi:10.48550/ARXIV.2203.01884

11. Petti AA, Williams SR, Miller CA, Fiddes IT, Srivatsan SN, Chen DY, et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. Nat Commun. 2019;10: 3660.

12. van Galen P, Hovestadt V, Wadsworth MH Ii, Hughes TK, Griffin GK, Battaglia S, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell. 2019;176: 1265–1281.e24.

13. Nam AS, Kim K-T, Chaligne R, Izzo F, Ang C, Taylor J, et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. Nature. 2019;571: 355–360.

14. Tansey W, Veitch V, Zhang H, Rabadan R, Blei DM. The holdout randomization test for feature selection in black box models. J Comput Graph Stat. 2022;31: 151–162.

15. Sirenko M, Lee S, Sun Z, Chaligne R, Asimomitis G, Brierley CK, et al. Deconvoluting clonal and cellular architecture in *IDH*-mutant acute myeloid leukemia. Blood. 2023;142: 1591–1591.

16. Timothy Tickle, Itay Tirosh, Christophe Georgescu, Maxwell Brown, Brian Haas. infercnv. Bioconductor; 2019. doi:10.18129/B9.BIOC.INFERCNV

17. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049.

18. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19: 15.

19. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20: 296.

20. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177: 1888–1902.e21.

21. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene Set Knowledge Discovery with Enrichr. Curr Protoc. 2021;1: e90.

22. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE; 2018. doi:10.1109/dsaa.2018.00018

23. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, et al. Captum: A unified and generic model interpretability library for PyTorch. 2020. doi:10.48550/ARXIV.2009.07896

24. Dong M, Blobe GC. Role of transforming growth factor-beta in hematologic malignancies. Blood. 2006;107: 4589–4596.

25. Wu Y, Su M, Zhang S, Cheng Y, Liao XY, Lin BY, et al. Abnormal expression of TGF-beta type II receptor isoforms contributes to acute myeloid leukemia. Oncotarget. 2017;8: 10037–10049.

26. Habbel J, Arnold L, Chen Y, Möllmann M, Bruderek K, Brandau S, et al. Inflammation-driven activation of JAK/STAT signaling reversibly accelerates acute myeloid leukemia in vitro. Blood Adv. 2020;4: 3000–3010.

27. Zhong F-M, Yao F-Y, Liu J, Zhang H-B, Li M-Y, Jiang J-Y, et al. Inflammatory response mediates cross-talk with immune function and reveals clinical features in acute myeloid leukemia. Biosci Rep. 2022;42. doi:10.1042/BSR20220647

28. Braun T, Carvalho G, Fabre C, Grosjean J, Fenaux P, Kroemer G. Targeting NF-kappaB in hematologic malignancies. Cell Death Differ. 2006;13: 748–758.

29. Zhou J, Ching YQ, Chng W-J. Aberrant nuclear factor-kappa B activity in acute myeloid leukemia: from molecular pathogenesis to therapeutic target. Oncotarget. 2015;6: 5490–5500.

30. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. 2019. doi:10.48550/ARXIV.1912.01703

31. Li L, Jamieson K, Rostamizadeh A, Gonina E, Hardt M, Recht B, et al. A system for massively parallel hyperparameter tuning. 2018. doi:10.48550/ARXIV.1810.05934

32. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023;39. doi:10.1093/bioinformatics/btac757

Chapter 4

Concluding remarks

4.1. Conclusion

Myeloid neoplasms are a complex set of prevalent and clonal hematologic malignancies characterized by genetic and phenotypic heterogeneity. Despite significant advances in the elucidation of the gene mutations that are frequently acquired in myeloid neoplasms, our understanding of the respective mechanisms, whereby these mutations cause disease pathogenesis, remains largely incomplete. To this end, advances in the development of patient relevant models of disease biology coupled to the deployment of high-throughput single and multi-omic laboratory assays set out to link established drivers of disease biology to specific molecular phenotypes. The scale and complexity of the data generated in research studies of MNs require the design and development of computational frameworks tailored to analyze such high-dimensional datasets. Thorough data analysis and interpretation of omic-based studies rely on the integration of multiple data views. This integration can be achieved either through the process of fusing data views towards facing supervised or unsupervised tasks (e.g. prediction, classification, clustering) or the process of interconnecting them towards studying the cross-talk between them.

This thesis develops analytical strategies based on multi-view data integration. We deploy data fusion and interconnection concepts to develop analytical frameworks within and across different omics views. We investigate the effects of the *SF3B1*^{K700E} mutation in the molecular landscape of MDS (Chapter 2) and capture prominent genotype-phenotype associations in *IDH1/2*-mutated AML (Chapter 3). The first study performs an interpretable multi-stage fusion of splicing, transcript usage and gene expression and results in a splicing signature that can accurately predict the *SF3B1* mutational status. Concurrently, the work presented here provides a comprehensive representation of the chromatin accessibility landscape and among others, nominates transcriptional programs with putative roles in MDS disease biology. In Chapter 3 (*IDH1/2* AML), we show that deep learning approaches applied on single cell gene expression and genotypic data, have the capacity to effectively predict the malignant status of cells and, importantly, the status of subclonal genomic abnormalities such as *NRAS*. Overall, the analytical frameworks presented herein demonstrate that the deployment of multi-view data integration concepts for the mining of bulk and single cell sequencing data in myeloid neoplasms, leads to a systematic and detailed profiling of molecular landscapes and enhances our ability to study genotype-phenotype relationships. The derived outcomes show that these approaches offer

the opportunity to establish links between diverse data views (e.g. links between splicing and transcript usage in Chapter 2 or effect of *NRAS* mutation on gene expression in Chapter 3) and characterize key signals related to disease biology.

The computational approaches of this thesis can be applied in extended datasets across MNs and other cancer indications. In a broader perspective, the rationale used for the integration of different data views from bulk RNA-seq data in Chapter 2, can be applied to other studies investigating the role of splicing factor mutations across relevant signals that can be quantified by transcriptome sequencing (expression, splicing, transcript usage). Given that mutations in splicing genes are not restricted to MDS, deploying the proposed multi-stage fusion framework in other diseases can enhance tissue-specific gene target prioritization and offer a thorough understanding of the transcriptomic repertoire through the derivation of signals from splicing analyses, transcript usage and gene expression. Additionally, concurrent bulk RNA and ATAC-seq in larger cohorts may enable the use of meta-dimensional fusion approaches that can jointly model patterns from the transcriptome and chromatin accessibility.

The supervised deep learning approaches of Chapter 3 can be deployed either to expand the genotype-phenotype associations derived herein in the context of MNs or to study the interconnection between single cell transcriptomes and genotypes in other translational studies. For instance, in *IDH1/2* AML, future efforts may focus on investigating links between the transcriptomic profiles and the genotypes of *IDH1* vs *IDH2* mutated cells. Other studies in oncology, especially in cancer indications with the presence of different genetic clones, can use matched transcriptomic and genomic data within classification or prediction tasks to explore if and how the acquired mutations and their clonality are reflected in the single cell gene expression profiles. We note that the approaches developed herein can be also used to study the associations between single cell genomic and chromatin accessibility profiles, as these architectures are compatible with the use of single cell ATAC-seq data as input instead of gene expression. However, a prerequisite for the efficient performance of the presented deep learning models is the presence of well-annotated and extensive data sets. Given though the challenges and limitations (e.g. cost, labor-heavy work) in aggregating larger sample sets and improving technical procedures upon the data generation process, future initiatives may focus on overcoming integration issues across single cell data (e.g. scRNA-seq) from different sources and technology protocols. This will enable the proposed frameworks to be easily applicable, without extensive retraining or calibration, on unseen patients in a wide research or clinical scale.

The fact that MNs present a phenotypic continuum of malignancies that share genomic abnormalities and treatment strategies, motivates the use of multi-view integration approaches at larger population scales on patients across the whole spectrum of the disease (MNs). Beyond bulk and single cell sequencing data, future incorporation of further data modalities that are commonly ascertained at diagnosis (such as digital pathology and immunophenotyping) into data integration strategies may have the power to reveal patient subgroups with molecular resemblance irrespective of their clinical annotations, and also unravel associations between genotypes and other data profiles.

It is worth highlighting though, that the addition of extra modalities needs to be coupled with the thoughtful formulation of scientific questions, the careful selection of cohorts and the mindful design of experimental

processes. Based on the experience acquired through the training process of this journey, I believe that the generation of data with minimum technical noise plus the collection and annotation of larger sample sizes are the foundation for data-driven research and provide the power to the downstream models to exploit the complementarity of the measurements and capture solid patterns between views. Moreover, from a computational perspective, considering the common challenges in processing and mining data from emerging omic technologies and other modalities, I note that efforts for consensus computational guidelines and literacy amongst research initiatives will be very beneficial for the advancement of translational research and will set the ground for the establishment of applicable pipelines with clinical utility. Given the multi-faceted nature of myeloid neoplasia, the collection of high quality multi-view datasets together with integration strategies and collaborative efforts from physicians, engineers and computational scientists are pivotal for the identification of the relevant molecular biomarkers and their adoption in clinical practice.

4.2. Data and code availability

A github repository containing the code used in generating the figures and the analysis results of Chapter 2 is available at the Papaemmanuil lab github page (https://github.com/papaemmelab/MDS_SF3B1_iPSC). The data used for this project (Chapter 2) are deposited in the Gene Expression Omnibus (GEO) under the accession number GSE184246. For Chapter 3, a github repository containing the models' source code is available at the Papaemmanuil lab github page (https://github.com/papaemmelab/Asimomitis_ACM_2023). The single cell dataframes used for the training, validation and testing of the models will become available upon publication of the Sirenko et al.