Dissertation

for obtaining the doctoral degree of the

Combined Faculty of
Mathematics, Engineering, and Natural Sciences
of the Ruprecht-Karls-University Heidelberg

Presented by

Carolin Andresen, M.Sc.
born in Eutin, Germany

Oral examination: 07.11.2022

Multi-omics analysis of DNMT3A- and NPM1-mutated acute myeloid leukemia

Referees:

Prof. Dr. Andreas Trumpp

Dr. Dr. med. Daniel Hübschmann

I. Abstract

Acute myeloid leukemia (AML) represents a genetically heterogeneous group of aggressive myeloid malignancies arising from clonal expansion of aberrant, myeloid-primed hematopoietic stem or progenitor cells. Intensive chemotherapy efficiently targets proliferating blasts and achieves remission in the majority of patients. However, most patients relapse, likely due to persisting, slowly proliferating leukemic stem cells (LSCs). A novel flow cytometry sorting strategy was recently developed in-house to enrich five different leukemic populations, two of them enriched for LSCs ($GPR56^{+}NKG2DL^{-}$). This strategy was applied to a genetically harmonized DNMT3A and NPM1 double-mutant AML cohort. Despite identical driver mutations, one group presented with early relapse (ER) while the other achieved long-term remission (LTR).

Multi-omics profiling (RNA-seq, DNA methylation, and genetic information) allowed me to deeply characterize these sorted leukemic populations and identify biological processes associated with ER. This analysis confirmed xenotransplantation experiments and demonstrated that the LSC-enriched populations exhibited indeed more stem-like characteristics. Still, LSC-enriched populations showed a higher cell cycle activity compared to non-engrafting, more differentiated AML populations. The LSC-enriched populations were transcriptionally similar, but the $CD34^{+}$ population retained also healthy hematopoietic stem cells (HSCs) while the $CD34^{-}$ population contained exclusively leukemic (stem) cells. This was particularly reflected by the distinct mutant allele frequencies of the DNMT3A- and NPM1-mutations. By analyzing the LSC-enriched populations, I demonstrated a higher transcriptomic instability in ER LSCs compared to LTR LSCs that may be initiated by increasing hypomethylation associated with an earlier onset of the DNMT3A mutation. Moreover, ER LSCs exhibited a more stem-like phenotype, characterized by higher activity of mitochondrial oxidative phosphorylation compared to LTR LSCs, which presented enhanced glycolytic activity instead. The difference in energy metabolism was partially confirmed by untargeted metabolomics analyses. In a technical development project, I also implemented an interactive R shiny app (MetaboExtract) and an R package (MetAlyzer) to infer suitable extraction protocols for metabolomics studies. In addition, I trained an outcome prediction expression signature to stratify patients based on their risk of relapse and hence long-term chemotherapy sensitivity. This signature was highly predictive in different AML cohorts and was able to stratify AML patients with poor and more favorable overall survival. In summary, my work revealed biological mechanisms associated with an early relapse in LSC-enriched AML populations and generated a novel outcome prediction signature to stratify patients.

II. Zusammenfassung

Die akute myeloische Leukämie (AML) stellt eine Gruppe genetisch heterogener, aggressiver Erkrankungen der myeloiden Hämatopoese, die durch klonale Expansion maligner hämatopoetischer Stammzellen (HSC) oder myeloischer Vorläuferzellen entstehen. Eine intensive Chemotherapie zielt effizient auf die proliferierenden Blasten ab, sodass bei der Mehrheit der Patienten eine Remission erreicht werden kann. Die meisten Patienten erleiden jedoch einen Rückfall, der wahrscheinlich auf persistierende, langsam proliferierende leukämische Stammzellen (LSCs) zurückzuführen ist. Eine neuartige FACS-Strategie zur Anreicherung fünf verschiedener leukämischer Populationen, einschließlich zweier LSC-angereicherter Populationen ($GPR56^{+}NKG2DL^{-}$) wurde kürzlich im Labor entwickelt. Diese wurde auf eine genetisch harmonisierte AML-Kohorte angewendet, deren Proben ausnahmslos Mutationen in den beiden Genen *DNMT3A* und *NPM1* tragen. Trotz identischer Treibermutationen erlitt eine Gruppe einen frühen Rückfall (ER), während die andere Gruppe eine langfristige Remission (LTR) erreichte.

Durch Multi-omics-Profiling (RNA-seq, Auslesen von DNA-Methylierungs- und genetischer Informationen) konnte ich die angereicherten leukämischen Populationen eingehend charakterisieren und jene biologischen Prozesse identifizieren, welche einen frühen Rückfall begünstigen. Diese umfassende Analyse bestätigte vorangegangene Xenotransplantations-experimente und zeigte, dass die LSC-angereicherten Populationen HSCs tatsächlich ähneln, aber dennoch eine höhere Zellzyklusaktivität aufweisen als die nicht transplantierten, stärker differenzierten AML Populationen. Die LSC-angereicherten Populationen waren transkriptionell ähnlich, jedoch umfasste die $CD34^{+}$-Subfraktion weiterhin gesunde HSCs, während die $CD34^{-}$-Subfraktion ausschließlich leukämische (Stamm-)Zellen enthielt. Dies spiegelte sich insbesondere in den unterschiedlichen Allelfrequenzen der DNMT3A- und NPM1-Mutationen wider. Durch die Analyse der LSC-angereicherten Populationen konnte ich zeigen, dass ER LSCs im Vergleich zu LTR LSCs eine höhere transkriptomische Instabilität besaßen, die möglicherweise eine stärkere Hypomethylierung durch ein früheres Auftreten der *DNMT3A*-Mutation ausgelöst wurde. Zudem wiesen die ER LSCs einen Stammzell-ähnlicheren Phänotyp auf. Dieser war gekennzeichnet durch eine höhere Aktivität der mitochondrialen oxidativen Phosphorylierung im Vergleich zu den LTR LSCs, die eine erhöhte glykolytische Aktivität aufwiesen. Der unterschiedliche Energiestoffwechsel konnte teilweise

durch die Anwendung von untargeted Metabolomik bestätigt werden. Im Rahmen eines technischen Projekts habe ich zudem eine interaktive, R Shiny App (MetaboExtract) entwickelt sowie ein R-Paket (MetAlyzer) implementiert, um Nutzern die interaktive Identifikation geeigneter Extraktionsprotokolle für Metabolomics-Studien zu ermöglichen. Darüber hinaus konnte ich eine Gensignatur trainieren, um Patienten anhand ihres Rückfallrisikos und damit ihrer langfristigen Sensitivität gegenüber Chemotherapie zu stratifizieren. Diese Signatur zeigte in verschiedenen unabhängigen AML-Kohorten eine hohe Vorhersagekraft und war in der Lage, AML-Patienten mit schlechtem und günstigem Gesamtüberleben zu stratifizieren.

Zusammenfassend zeigt meine Arbeit biologische Mechanismen auf, die mit einem frühen Rückfall in LSC-angereicherten AML-Populationen assoziiert sind, und präsentiert eine neue Genexpressions-Signatur zur Stratifizierung von Patienten anhand ihrer voraussichtlichen Überlebenswahrscheinlichkeit.

Contents

1 Introduction

1.1 Acute Myeloid Leukemia

Acute myeloid leukemia (AML) is the most common acute leukemia in adults with an incidence of 4.3 cases per 100,000 people in the US.[1] The incidence increases with age leading to a median age at diagnosis of about 65 years.[2] Intensive research has allowed an increasingly precise classification of AML based on molecular and cytogenetic abnormalities and treatment has been improved over the past decades.[3] Still, AML has high mortality rates. The estimated 5-year survival is about 25% with a significantly poorer prognosis in the elderly.[4] While 35-40% of patients younger than 60 years are cured, only 5-15% of older patients present long-term therapeutic success.[3] The origin of AML is heterogeneous; in some patients, exposure to DNA-damaging agents or chemotherapy is the suspected cause of the disease or AML arises as a progression of other hematologic malignancies. However, most cases appear *de novo* without clear etiology.[1,5] With intensive induction therapy, complete remission can be achieved in 60-85% of cases, still, most patients ultimately relapse.[1] One main explanation is the incomplete eradication of leukemic stem cells (LSCs). These cells are either primarily resistant to standard therapy or develop secondary resistance mechanisms and have an ability to self-renew and reinitiate the disease.[6–10]

1.2 The origin of AML

1.2.1 Hematopoiesis

Blood has a multitude of different functions, from oxygen transport and supply of nutrients to the various tasks of the immune system.[11] Most of the mature blood cells are relatively short-lived and constant replenishment is needed to ensure homeostasis and functionality of the system.[12] Hematopoiesis is hierarchically organized with hematopoietic stem cells (HSCs) at the apex (Figure 1). HSCs are multipotent cells that differentiate into progenitor cells by an asymmetric division which produces one daughter progenitor cell while maintaining one stem cell.[13] By differentiation, progenitors become increasingly specialized and finally give rise to mature functional blood cells which show a great morphological and functional diversity.[14] The hierarchical tree is divided into a lymphoid branch which differentiates into, e.g., T, B, and natural killer (NK) cells while the erythrocyte-myeloid lineage gives rise to, e.g., monocytes, megakaryocytes, and erythrocytes.[15] The cells of the hematopoietic system have been intensively studied and well characterized by phenotypic marker genes presented on

their surface. These markers can be used to isolate specific cells via fluorescence-activated cell sorting (FACS). For example, hematopoietic stem and progenitor cells (HSPCs) are characterized by a $CD34^+$ immunophenotype.[16,17] The tree presented in Figure 1 is a simplified model of hematopoiesis. Indeed, the hierarchy is more plastic and complex with shortcuts and return paths.[18] Furthermore, with the advent of single-cell technologies, a rather continuous than discrete differentiation model was proposed.[19]

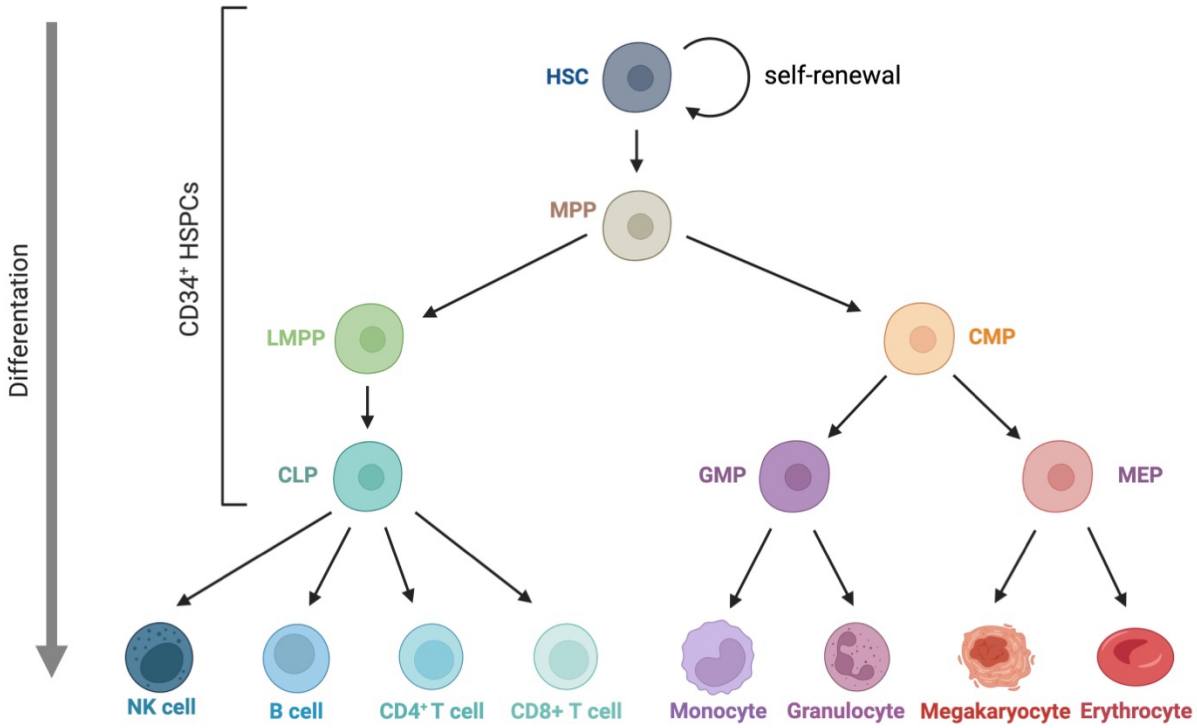

Figure 1: Simplified hematopoietic hierarchy. The multipotent HSC is located at the apex of the hierarchy giving rise to different progenitor cells which progressively differentiate into the lymphoid lineage and the erythrocyte-myeloid lineage. In healthy hematopoiesis, stem and progenitor cells (HSPCs) are characterized by a $CD34^+$ immunophenotype. HSC: hematopoietic stem cells; MPP: multipotent progenitor; LMPP: lymphoid-primed multipotent progenitor; CMP: common myeloid progenitor; CLP: common lymphoid progenitor; GMP: granulocyte-macrophage progenitor; MEP: megakaryocyte-erythroid progenitor; NK: natural killer cells. (Adapted from Corces et al.[20])

HSCs reside in the bone marrow where hematopoiesis and maturation of blood cells mainly take place. The bone marrow provides a physically and molecularly protective microenvironment for these crucial cells and has been shown to be essential for HSC function.[21,22] The low abundant HSCs population is characterized by their ability to self-renew,

allowing these cells to generate new, identical daughter HSCs after cell division.[23] Based on the immunophenotype of these cells, different FACS-sorting strategies were proposed and it became apparent that HSCs themselves are a heterogeneous group of cells.[24–28] Even though some HSCs cycle to replenish the progenitor pool, most of them are in a quiescent state.[28,29] So-called dormant HSCs have been characterized as the most potent HSCs by long-term label-retaining assays and serial transplantation.[28,29] These cells divide only a few times per lifetime, which protects them from stress and the risk of accumulation of mutations.[28] However, in response to excessive stress and injury signals (e.g. induced by infections or chemotherapy) dormant HSCs also get activated to maintain the essential functions and homeostasis of the hematopoietic system.[28,30]

1.2.2 Leukemogenesis and leukemic stem cells

AML is a heterogeneous hematopoietic malignancy that arises from clonal expansion of genetically aberrant HSCs or myeloid-committed progenitors as cells of origin (Figure 2).[8,31–33] Leukemogenesis of AML is based on multiple causes, steps, and pathways.[34] Over their lifespan, individuals accumulate mutations that are linked to the division and proliferation of cells also in the slow-cycling HSPCs.[29] Hence, mutations in HSPCs increase with age. AML can arise *de novo* or develop from pre-leukemic conditions such as myelodysplastic or myeloproliferative disorders.[34,35] Clonal hematopoiesis of indeterminate potential (CHIP) describes the clonal expansion of mutated HSCs and also has increasing prevalence with increasing age. Even though it does not necessarily entail clinically relevant consequences, it increases the risk of developing AML if pre-leukemic HSCs acquire further mutations.[36,37] Early driver mutations are often epigenetic regulators (e.g., *DNMT3A* or *TET2*) that prime the cells before additional, later mutations initiate a leukemia (Figure 2).[38,39]

Similar to the healthy hematopoietic system, a hierarchical organization has been observed for AML.[40–42] In AML, LSCs reside at the apex of the hierarchic tree, but instead of differentiating into functional cells, a large number of immature leukemic blasts is produced.[43] Accumulation of these non-functional blood cells leads to severe clinical symptoms attributable to the loss of mature blood cells.[33,44] LSCs have been characterized by their self-renewal capacity, relative quiescence, and resistance to apoptosis. Additionally, LSCs are often more resistant to conventional chemotherapy which rather targets the more proliferative leukemic progenitors and blasts, but also increased drug efflux mechanisms have been described for LSCs.[7] Hence, while chemotherapy is often successful to eliminate the

leukemic bulk, reflected by a complete remission, LSCs survive and initiate a relapse, the cells of which are often even more resistant to therapeutic interventions (Figure 2). Consequently, based on this cancer stem cell model, eradication of LSCs is essential for long-term remission or cure.[6,45,46] Studies have shown that LSCs share similar characteristics with HSCs, however it is crucial to study these leukemic cells in order to identify therapies specifically targeting LSCs.[47,48]

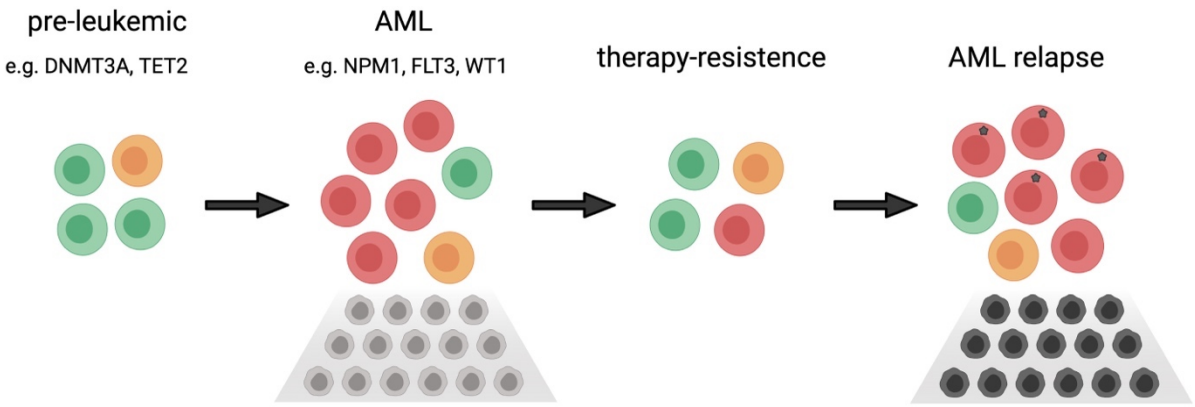

Figure 2: Schematic development, therapy, and relapse of *de novo* AML. A healthy HSPC (green) acquires a pre-leukemic (orange) phenotype (e.g., mutation of *DNMT3A*). Subsequent mutations (e.g., *NPM1*) give rise to AML. LSCs (red) differentiate into non-functional leukemic blasts (grey). Therapy eradicates the blast fraction but resistant LSCs persist and initiate the relapse. A black asterisk indicates the potential development of resistance mechanisms or the proliferation of a dominant (potentially therapy-induced) subclone.

Like HSCs, LSCs are a rare population with an estimated number of 1 LSC per 1 million leukemic blasts. Experimentally, LSCs have been defined by their capacity to engraft in xenotransplantation assays which can be propagated multiple times in serial transplants, as well as their ability to produce leukemic progeny that lacks the ability to engraft.[49,50] Using this definition, Ng et al. could show that LSCs, defined by a 17-gene stemness score (LSC17), are indicative for the survival of patients.[51] In addition, many studies aimed to identify markers that immunophenotypically characterize LSCs to ultimately target or isolate and study these cells.[52–56]

1.2.3 The genetic landscape of AML

With advancements in next-generation sequencing (NGS) technologies, various studies have been set up aiming to investigate the genomic landscape of AML in more detail based on whole genome sequencing (WGS).[57,58] As compared to other adult cancers, AML genomes have a relatively low number of mutations (13 on average), some of which have a high recurrence.[59] The most commonly mutated genes are *FLT3* (˜28%), *NPM1* (˜27%), *DNMT3A* (˜26%), *IDH1* or *IDH2* (˜20%), and *NRAS* or *KRAS* (˜12%).[59] As shown in table Table 1, mutations often affect similar biological processes.

Table 1: Frequently mutated genes in AML by functional categories. (Adapted from Döhner et al.[3])

Functional Catagory	**Gene (Frequency [%])**
Nucleophosmin	*NPM1* (25-35)
Tumor suppressor genes	*TP53* (~8), *PTEN*, *NRAS* (~12)[59], *KRAS* (~12)[59]
DNA methylation	*DNMT3A* (18-22), *IDH1* (7-14), *IDH2* (8-19), *TET2* (7-25)
Splicosome complex	*SRSF2*, *SRF3B*, *U2AF1*, *ZRSR2*
Cohesin complex	*STAG2*, *RAD21*
Transcription factor fusions	*RUNX1* (5-15), *RUNX2*
Activated signaling	*FLT3*-ITD (~20) [*RAS-RAF*, *JAK-STAT*, *PI3K-AKT*]
Chromatin modifications	*DOT1L*, *KMT2A* (~5), *MLLT3*, *EZH2*, *ASXL1* (5-17)
Others	*CEBPA* (6-10), *KIT* (<5)

Analyses of the mutational landscape also revealed clonal heterogeneity at diagnosis; and clonal evolution observed at relapse is suspected to contribute to therapy resistance.[59,60] Based on studies analyzing the clonal evolution, it is assumed that mutations in genes involved in epigenetic regulation often occur early in AML progression and contribute to a pre-leukemic phenotype.[32,39] One of the most frequently mutated epigenetic modifiers is DNA (cytosine-5)-Methyltransferase 3A (*DNMT3A*). This gene is involved in the establishment of *de novo* methylation.[3,61] Mutations at various positions in the gene have been described. Depending on the position of mutation, different functional changes and entailed altered methylation patterns have been observed.[62,63] However, *DNMT3A* is most frequently mutated at protein position p.R882 which probably causes a loss of function.[64,65] The loss of *Dnmt3a* immortalizes HSCs by hypomethylation of regions associated with self-renewal genes.[66] Therefore, Dnmt3a loss of function skews HSC division towards self-renewal and to an increasing outcompetition against normal HSCs, particularly in older individuals.[66] In

general, *DNMT3A* mutations have been associated with poor overall survival.[67] However, this might depend on the mutation, and a worse outcome was observed only for patients with R882 mutations but not with non-R882 mutations.[68] Mutations of *DNMT3A* frequently co-occur with mutations of *NPM1* and *FLT3*. 60-80% of *DNMT3*A-mutated cases also harbor an *NPM1* mutation and about 30% are triple-mutant for *DNMT3A*, *NPM1,* and *FLT3*.[65,69]

Nucleophosmin (*NPM1*) is one of the most frequent driver mutations in AML.[70] Wild type *NPM1* is usually located in the nucleolus but can also be found in the cytoplasm and nucleoplasma.[71] Many functions of the gene have been discussed. Its involvement in cellular processes can be summarized into four major areas: ribosome biogenesis, p53-dependent stress response, genomic stability (e.g., centrosome duplication), and modulation of growth-suppressive pathways.[72] Observed pathogenic mutations in *NPM1* cause a frameshift in the C-terminus, which lead to a stronger nuclear export and therefore aberrant cytoplasmic localization of the protein. These mutations are also referred to as *NPM1*c.[70,72,73] These mutations are always heterozygous, which in turn is in line with embryonic lethality upon double knockout *Npm1* mice.[74] Even though *NPM1* has been identified as a frequent driver, the exact mechanism in leukemogenesis remains elusive. Since *NPM1* has been linked to various cellular processes, both loss and gain of gene function have been proposed to drive AML development. In general, *NPM1* mutations are associated with a favorable prognosis depending on co-occurring events.[75]

Apart from mutations, cytogenetic alterations are frequently observed in AML patients and have been used to define genomic subgroups.[76] For example, the karyotype is a strong prognostic factor defining a subgroup of so-called complex karyotype AML with a highly unfavorable prognosis.[77]

1.3 Clinical aspects of a complex disease

1.3.1 Classification and prognostic factors

Historically, AML was classified based on the morphologic phenotype by the French-American-British (FAB) groups. These groups are based on myeloid lineage involvement and differentiation presented in histochemically stained blood smears.[44] More recently, the "*World Health Organization (WHO) Classification of Tumours of Haematopoietic and Lymphoid Tissues*" classified AML based on molecular genetic lesions. This classification,

updated in 2016, defines six main AML-related groups, including "AML with recurrent genetic abnormalities".[31]

The prognosis of the disease depends primarily on the combination of genetic drivers, molecular and cytogenetic risk, as well as patient-associated factors such as age.[78] The ELN (European LeukemiaNet) stratifies AML into three subgroups (favorable, intermediate, adverse) based on molecular genetics and cytogenetic alterations (Table 2).

Table 2: Risk stratification according to 2017 ELN recommendations. Classification based on molecular genetics and cytogenetic alterations. (Adapted from Döhner et al.[79])

Risk category	Genetic abnormality
Favorable	t(8;21)(q22;q22.1); *RUNX1-RUNX1T1*
	inv(16)(p13.1;q22) or t(16;16)(p13.1;q22); *CBFB-MYH11*
	Mutated *NPM1* without *FLT3*-ITD or with *FLT3*-ITD^{low}
	Biallelically mutated *CEBPA*
Intermediate	Mutated *NPM1* and *FLT3*-ITD^{high}
	Wild type *NPM1* without *FLT3*-ITD or with *FLT3*-ITD^{low} (without adverse-risk genetic lesions)
	t(9;11)(p21.3;q23.3); *MLLT3-KMT2A*
	Cytogenetic abnormalities not classified as favorable or adverse
Adverse	t(6;9)(p23;q34.1); *DEK-NUP214*
	t(v;11q23.3); *KMT2A* rearranged
	t(9;22)(q34.1;q11.2); *BCR-ABL1*
	inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2*,*MECOM(EVI1)*
	−5 or del(5q); −7; −17/abn(17p)
	Complex karyotype, monosomal karyotype
	Wild type *NPM1* and *FLT3*-ITD^{high}
	Mutated *RUNX1*
	Mutated *ASXL1*
	Mutated *TP53*

Besides patient age (60+ years), negative prognostic factors include, e.g., coexisting conditions and illnesses, high white blood cell count, and female sex. In clinical practice, particularly age, correlated with general health, is an important clinical stratification since older patients are often unable to receive intensive chemotherapy.[3,44] In addition, supportive care is essential to increase survival during intensive therapy and greatly influences patient outcome. This includes transfusion of red blood cells and platelets as well as antibiotic and anti-fungal treatment in immunocompromised patients.[1,80,81]

1.3.2 Clinical presentation and diagnosis

AML leads to symptoms derived from cytopenia of all normal mature hematopoietic cells. These symptoms are often unspecific and include fatigue, pallor, hemorrhage, and recurrent infections. Patients typically present with reduced mature hematopoietic cell counts, e.g., for erythrocytes and platelets. The primary diagnosis includes the morphological assessment of a peripheral blood smear stained with Wright-Giemsa, however, a definitive diagnosis requires bone marrow examination and >20% infiltration with leukemic blasts. These leukemic myeloblasts display distinct morphological aberrations such as irregular nuclei, little cytoplasm, and presence of azurophilic granules called Auer bodies or rods. It is essential to make the correct diagnosis in order to identify appropriate therapeutic strategies. Hence, immunohistochemical examination and genetic evaluation require careful analysis to diagnose the AML subtype, and distinguish from similar diseases such as acute lymphoblastic leukemia (ALL) and myelodysplastic syndrome (MDS).[3,44,79]

1.3.3 Treatment

First-line treatment of AML has not substantially changed in the last decades and still consists of intensive chemotherapy divided into induction therapy and consolidation phase. This first phase aims to reach a complete remission in patients defined by a <5% bone marrow blast count, recovery to normal peripheral blood cell counts and absence of signs and symptoms.[3,44] In AML, the standard chemotherapy for patients in good physical shape follows the so-called “7+3” regimen consisting of seven days cytarabine treatment and then three days of treatment with an anthracycline (usually daunorubicin or idarubicin).[4] After complete remission, low numbers of leukemic cells are still present and lead to a relapse in almost all patients. Therefore, the consolidation phase is essential to reach long-term remission or even cure of the disease. This phase mainly depends on the diagnosis, risk stratification, and patient-related prognostic factors. Besides intensive chemotherapy, the consolidation might include allogeneic stem cell transplantation. However, only a minority of younger and fit patients are eligible for transplantation even if a donor is available.[1,3]

In recent years, particularly for unfit and relapsed patients, novel targeted therapies have been approved such as hypomethylating agents (decitabine, azacitidine) or IDH1 inhibitors (ivosidenib). More recently, the BCL2-inhibitor venetoclax demonstrated anti-leukemic efficacy and synergy with other therapeutic agents for AML.[82–84] Further studies proposed

that venetoclax in combination with azacitidine might specifically target LSCs.[85,86] More experimental approaches also aim to render dormant HSCs or LSCs sensitive to chemotherapeutic approaches.[87–89]

1.4 Studying AML and LSCs

In recent decades, technological advances have facilitated more sophisticated investigation at an increasing rate. Particularly, high-throughput technologies have enabled precise characterization of biological samples, also in AML research. For example, sequencing methods allowed the identification of common driver genes and investigation of the genetic, epigenetic and transcriptional landscape of the disease.[57,59,90] Laboratory techniques have also become more advanced and new software solutions have been developed to support the analysis of large data sets.

1.4.1 Marker genes

Characterization of the hematopoietic system has been driven by immunophenotyping of cells. Because they are mostly in liquid phase, blood cells are ideal for FACS analysis and in recent years a variety of surface markers have been established to enrich or isolate certain cell populations.[16] Large efforts have been made to identify and phenotypically characterize LSCs based on the cell surface phenotype. Since complete eradication of these cells with disease-initiating capacity may ultimately prevent relapse and therapy resistance, identification and isolation of LSCs is of highest importance.[52–56] Many studies applied FACS-sorting strategies to enrich for LSCs derived from experience gained in the healthy hematopoietic system ($CD34^+CD38^-$).[42,91,92] However, cell phenotypes show high variability between specimens and genetic subgroups, and markers used within healthy hematopoiesis cannot always be transferred to the diseased system.[93] Here, several markers have been suggested, e.g., CD123, CD47, TIM3, CD25, CD32, and ALDH. However, studies demonstrated high intra-patient heterogeneity. Some samples contained LSCs in non-LSC populations while other AML samples did not engraft at all.[52–56,93] Hence, enrichment for LSCs by cell surface phenotype must be validated by xenotransplantation assays, the defined standard method to prove LSC activity.[50]

Recently, two novel potential markers for LSCs were proposed that enrich for LSCs independent from the classic CD34 marker for healthy HSCPs. Papst et al. described that high expression of adhesion G Protein-Coupled Receptor 56 (GPR56, encoded by *ADGRG1*) is indicative of engraftment potential in both the $CD34^+$ and $CD34^-$ populations. In addition, they showed significant association of GRP56 expression with high-risk genetic AML subgroups and poor disease outcome.[94] The gene GPR56 is also part of the LSC17 signature associated with outcome and stemness.[51] This receptor is expressed in various tissues and best studied in the nervous system, but also in HSPCs.[95] In mice, GRP56 is expressed in HSPCs with decreasing levels upon differentiation. It is potentially involved in overcoming proliferative stress, but seemed to be dispensable in steady state since no significant changes regarding maintenance or migration of HSPC could be observed.[96] Other studies showed expression of GPR56 in cytotoxic lymphocytes, particularly in NK and different T cells.[97]

Moreover, Paczulla et al. identified ligands of the Killer Cell Lectin Like Receptor K1 (NKG2D, encoded by *KLRK1*) as a novel marker for LSCs. Cells negative for NKG2D ligands (MICA, MICB, ULBP1, ULBP2, ULBP5 and ULBP6) were isolated using a NKG2D-Fc chimeric protein. Even though some cell types in AML samples did express NKG2DL (NKG2D ligands), this was never the case for LSCs. Hence, $NKG2DL^-$ populations were identified as enriched for LSCs, again in both the $CD34^+$ and $CD34^-$ populations. The study also showed that negativity for NKG2D ligands facilitates immune evasion of the LSC population.[98] Of note, the receptor NKG2D is expressed by NK cells and a subset of T cells and acts as danger detector that mediates elimination of transformed or infected cells. Usually ligands of NKG2D are lowly expressed but can be induced by different pathways activated upon infection or during tumorigenensis.[99,100]

These novel phenotypic markers for LSCs are of particular interest when studying $CD34^-$ AMLs defined by CD34 present on <10% of all leukemic blasts. The low expression of CD34 is especially prevalent in NPM1-mutated AMLs and makes up about 25% of all AML cases.[93,101]

1.4.2 Methylation

DNA methylation is a crucial epigenetic modification, which allows changing the activity of a sequence element without changing the sequence itself and provides a mechanism to pass on gene expression patterns across cell divisions and to progeny cells.[102] A very important

fraction of human DNA methylation occurs in the context of CpG nucleotides (dinucleopair cytosine and guanosine), forming 5-methylcytosine by the addition of a methyl group (CH_3) to the nucleoside cytosine.[102] This modification is often associated with the silencing of genes by either directly interfering with the binding of regulatory proteins or the activation of chromatin remodeling into inactive heterochromatin.[102–104]

When averaging across various cell types, about 70% of CpGs in the human genome are methylated. However, there are CpG-rich clusters, so-called CpG islands, which are predominantly unmethylated. CpG islands are regulatory units in the DNA often located in the promoter regions of genes.[103,105] However, DNA methylation also occurs in other regions than CpG islands. These regions are often defined as "shore", "shelf", and "open sea" depending on the distance to CpG islands (Figure 3).[106]

Figure 3: Scheme of methylated regions and distances to CpG islands.

Particularly during cell development and differentiation, methylation is dynamic and essential for maintaining cellular identity.[107] Methylation patterns are maintained during replication by certain DNA methyltransferases (e.g., DNMT1), whereas *de novo* DNA methyltransferases (e.g., DNMT3A), modify unmethylated CpG nucleotides later in the development or differentiation.[102,104] In general, DNA methylation increases with differentiation from stem cells to more committed cells.[107] This pattern is also observed in lymphoid differentiation during hematopoiesis. An exception from the general rule is myeloid differentiation, in which DNA methylation generally decreases, although it is dynamically regulated at different stages.[108,109]

Aberrant DNA methylation in cancer has been widely described and focuses primarily on promoter hypermethylation causing the silencing of tumor suppressor genes.[104,110] In AML, global hypomethylation has been observed and associated with genomic instability.[111] In *DNMT3A*-R882H-mutated cells, CpG island hypomethylation has been described, while other *DNMT3A* mutations caused CpG island hypermethylation, potentially as a consequence of AML progression.[62]

1.4.3 Alternative Splicing

Splicing describes the crucial processing from precursor messenger RNA (mRNA) to mature mRNA by removal of non-coding intronic regions and subsequent fusion of coding exons before protein translation. This process is mostly carried out by the spliceosome, which is a tightly regulated machinery of small ribonucleoprotein complexes (snRNPs). The spliceosome recognizes 5' and 3' splice sites which are conserved *cis*-elements at exon-intron boundaries. While mRNA splicing removes non-coding introns, exonic regions may be alternatively retained or excluded from the final transcript (Figure 4). This process is referred to as "alternative splicing", leading to alternative isoforms of mRNA which can translate into proteins with distinct cellular functions.[112,113] Hence, alternative splicing greatly diversifies the human proteome.[114]

The regulation of alternative splicing is influenced by *cis*-regulatory regions in exons or introns which act as enhancers or silencers. These regions are targeted by *trans*-acting RNA-binding proteins (RBPs), heterogeneous nuclear ribonucleoproteins (hnRNPs), and serine and argine-rich proteins (SR proteins).[112,115] Additionally, the relevance of secondary structure of the precursor mRNA, sequence modifications and epigenetic changes such as DNA methylation for the regulation of alternative splicing have been investigated.[115–117]

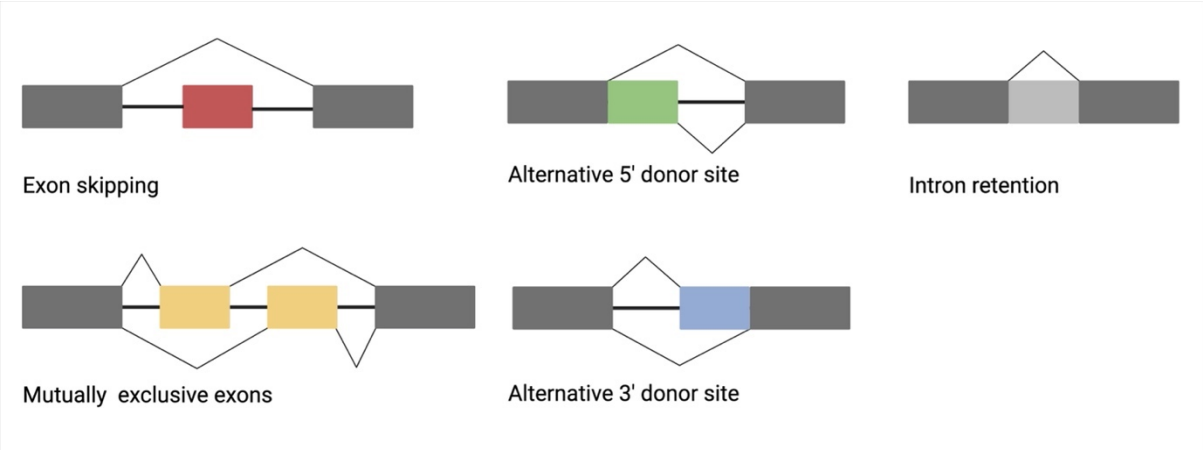

Figure 4: Different types of alternative splicing events covered by the rMATS software. Exons are represented by blocks. Introns are represented by thick lines. Thin lines indicate where splice sites are merged to build spliced mRNA. (Adapted from Shen et al.)[118]

Alternative splicing in AML has been described repeatedly. It acts through different mechanisms and could serve as a prognostic marker.[119–121] Splicing factors responsible for the

assembly of the core spliceosome such as Serine and Arginine-Rich Splicing Factor 2 (SFRS2) or Splicing Factor 3b Subunit 1 (SF3B1) are frequently mutated in AML (cf. Table 1).[122] Even though aberrantly spliced transcripts might be eliminated by RNA control mechanisms like nonsense-mediated decay, alternatively spliced transcripts are actually observed frequently in AML. Aberrant mRNA transcripts have also been described to correlate with transcriptomic instability, the latter being itself associated with *DNMT3A* mutations.[123–125] For example, alternative splicing of EZH2 has been described to lead to decay of this tumor suppressor and therefore to phenocopy loss-of-function mutations.[121,126] Another example is the generation of tumor-specific (neo)antigens by alternative splicing, which could act as targets for AML immunotherapy.[119] Additionally, multiple studies aimed to train predictive splicing signatures and proposed improved prognostic classification of AML.[120,127–130] Since alternative splicing leads to a complex rearrangement of the transcribed sequence, the identification of events and quantification of alternative isoforms is not trivial. Various methodological approaches have been proposed including exon-based, isoform-based, and event-based methods, while more recent approaches focus on the quantification of differential expression of splice junctions.[131–133]

1.4.4 Metabolomics

Metabolic products are the direct outcome of all cellular processes and thus represent the biological endpoint of the omics cascade. Metabolomics is an evolving field of research generating data that is complementary to genetic, transcriptomic or proteomic information [134]. Analysis of metabolic data has been used for numerous applications, e.g., for biomarker or drug discovery in various areas, including cancer research.[135,136,145,146,137–144] Metabolites are molecules with very diverse structures and chemical properties. High-throughput quantification relies on mass spectrometry coupled to other technologies such as gas or liquid chromatography. Metabolomics can be grouped into untargeted and targeted approaches. [147,148] Untargeted analyses are suited for discovery-driven studies for hypothesis generation. This technique allows relative quantification and qualitative identification of spectra based on comparison to libraries for metabolite structures. In contrast, targeted approaches focus on the analysis of known metabolites by comparison to reference standards. This analysis is often more hypothesis-driven by absolute quantification of specific metabolites of interest.[149]

In the past years, commercial kits became available which measure a relatively large number of metabolites within relevant pathways and allow absolute quantification by providing internal standards and calibrants, such as the MxP® Quant 500 kit (Biocrates). Metabolomic measurements are well-established for analysis of body fluids such as blood plasma, cerebrospinal fluid (CSF) or urine.[150–153] The liquid nature of these samples allows fast sample processing. However, the measurement of intracellular metabolites requires more complex protocols for pre-analytical extraction including tissue homogenization, removal of extracellular compounds, and lysis of cells. This leads to intracellular measurements being technically much more challenging since complex pre-processing might introduce technical variance. Beside the extraction itself, measurements of metabolites are challenging due to the fast turnover of metabolites, the need for normalization to sample cell number, comparability, reproducibility and extraction of molecules with heterogeneous chemical properties.[154,155]

Recent publications analyzing small sets of metabolites showed that the metabolome of AML is profoundly altered.[86,156,157] Particularly for LSCs metabolic vulnerabilities might offer potential therapeutic options. As outlined above, designing targeted therapies for LSCs is especially challenging since these cells are more quiescent and share characteristics with HSCs which must be spared from therapy.[158,159] It is therefore necessary to identify features which are differential between these two cell types. For example, differences in energy metabolism have been described. HSCs rather rely on glycolysis and low activity of oxidative phosphorylation and abundance of reactive oxygen species (ROS) have been observed, which also protects these quiescent cells from DNA damage. Only the differentiation into progenitor cells and accordingly higher energy demand led to a metabolic switch towards oxidative phosphorylation.[29,160,161] LSCs show a similar metabolic profile as HSCs, however, these cells seem to rely on oxidative phosphorylation for energy production. Various studies have shown the energy metabolism might be a selective target for LSCs.[85,162,163]

1.5 The leukemic hallmarks

Tumorigenesis is a multistep process and often similar functional aberrations are involved. Hanahan & Weinberg conceptualize these underlying principles as the hallmarks of cancer. The most recent version includes 14 hallmarks such as evading tumor suppressors, deregulation of cellular metabolism, avoiding immune destruction, and sustaining

proliferative signaling.[164] These mechanisms have also been described to be involved in the leukemogenesis of AML.

A proliferative advantage of AML cells seems to be often induced by constitutive activation of signaling pathways involved in growth, proliferation, survival, and proliferation. For example, PI3K/AKT/mTOR or JAK-STAT signaling is frequently hyperactivated in AML (cf. Table 1).[158,165] Analogously proliferation might be induced by the evasion of growth suppressors. The antiproliferative effects of Transforming Growth Factor β (TGFβ) signaling is well studied, however, ambiguous role at different cancer stages have been described.[166,167] Enhanced proliferation implies an increased cell cycle activity. The cell cycle is divided into different phases that indicate whether a cell is quiescent (G_0), growing (G_1 and G_2), in the DNA replication phase (S) or dividing (M).[168] In cancer, the cell cycle is often deregulated by a mutation-induced loss-of-function of tumor suppressor genes (e.g. *TP53*, cf. Table 1). These genes are key regulators governing the decision to either proliferate or activate apoptotic programs, e.g., when sensing DNA damage or abnormal stress.[167,169]

Growing and proliferating cells have been shown to adjust their energy metabolism. Under aerobic conditions, normal cells mostly metabolize glucose via the tricarboxylic acid (TCA) cycle and subsequent oxidative phosphorylation for efficient production of adenosine triphosphate (ATP) via the electron transport chain. Cancer cells, but also proliferating cells, switch the energy production towards aerobic glycolysis where influx of glucose is increased, and glucose is metabolized to lactate. While this alternative metabolic route is less efficient to generate ATP, intermediates of increased glycolysis can be fed into various biosynthetic pathways to generate molecules needed for proliferation. This function is also referred to as the Warburg effect and describes this characteristic of cancer cell energy metabolism (Figure 5).[161,164,170–172] As described above, energy metabolism in AML is subtle as different cells within the AML hierarchy utilize different routes of energy metabolism to different extent, with specific characteristics for LSCs.

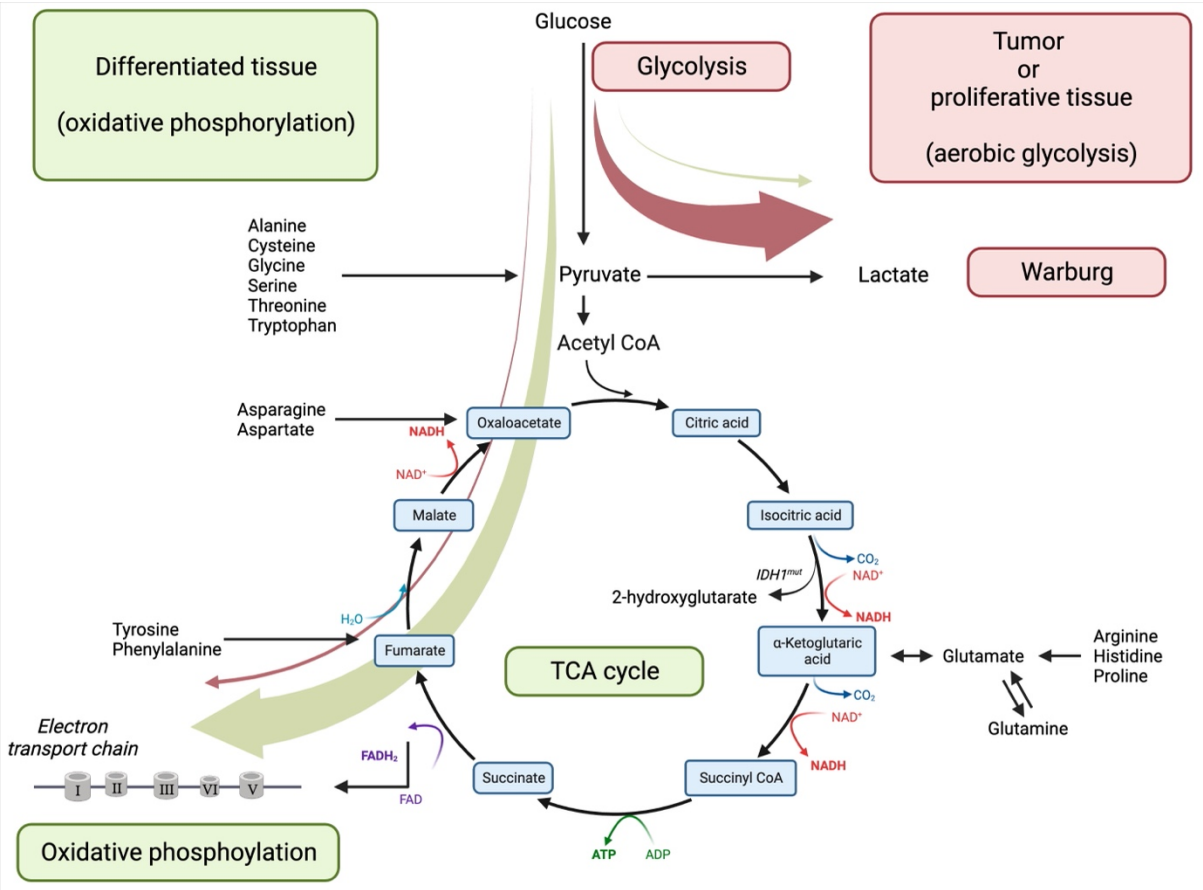

Figure 5: Energy metabolism in normal and tumor cells. In differentiated tissue, glucose is mainly metabolized via the TCA cycle and subsequent oxidative phosphorylation. In tumors and also proliferative tissue, cells consume high levels of glucose, which is metabolized mainly into lactate. Of note, the simplified scheme shows the energy metabolism under aerobic conditions. In differentiated tissues, pyruvate is also metabolized into lactated when oxygen is not present. In Isocitrate Dehydrogenase 1 (IDH1)-mutant cells, isocitric acid is converted into 2-hydroxyglutarate instead of α-ketoglutaric acid. NAD^+(H): nicotinamide adenine dinucleotide (oxidized/reduced); CO_2: carbon dioxide; ADP: adenosine diphosphate. (Adapted from Erdem et al. and Fadaka et al.)[161,171]

Beside mechanisms which sense cellular stress and damage such as tumor suppressor genes, also the immune system is a first line defense to detect and eliminate aberrant cells.[167] These mechanisms are impaired in tumor cells and their niche, often due to defects of immune checkpoints. A therapeutically very important immune checkpoint is the interaction of Programmed Cell Death Protein 1 (PD-1) and its ligand (PD-L1). The interaction inhibits T cell response as a mechanism to prevent autoimmunity in normal cells. Many cancer cells utilize this and other immune checkpoints to evade the immune system.[173,174]

Unlike in some solid cancers, in AML, inhibition of immune checkpoints showed limited response, likely due to the low overall number of mutations, a consecutive lack of neo-epitopes and thus a lack of actionable targets for the T cells. T cells recognize mutated cancer cells by the surface presentation of aberrant peptides, the neo-epitopes, via major histocompatibility complex (MHC) I. Hence, low mutational burden correlates with a low number of immunogenic antigens on AML cells.[119,175] While MHC-I molecules are presented by all nucleated cells, MHC-II molecule expression is restricted to antigen presenting cells (APCs). Interestingly, recent studies showed that HSPCs constitutively present antigens via MHC-II as an immunosurveillance mechanism and loss of antigen surface representation is associated with relapse in AML patients.[176,177] Taken together, during leukemogenesis, various aberrant processes have to act together to overcome cellular safety mechanisms and to induce abnormal, malignant proliferation.

1.6 SyTASC - Systems-based Therapy of AML Stem Cells

The SyTASC (Systems-based Therapy of AML Stem Cells) consortium has put together a cohort of adult AML patients with a homogeneous genetic background. All 38 samples harbor pre-leukemic mutations in *DNMT3A* (p.R882) and leukemic driver mutations in *NPM1* (p.W288fs*12). Even though almost identical driver mutations were identified across the cohort and patients showed complete remission after chemotherapy, some patients showed rapid relapse while others achieved long-term remission. The cohort was therefore stratified into two outcome groups; (i) early relapse (ER) patients that relapsed less than 6 months after treatment and (ii) long-term remission (LTR) patients (Figure 6).

As part of the SyTASC project, a novel sorting strategy was established by Dr. Nadia Correia to particularly study the LSC subpopulations and their differences between the two outcome groups as these disease-initiating cells are essential for the relapse of patients. The sorting strategy included the novel FACS markers NKG2DL and GPR56 as well as the traditional HSC marker CD34 to sort five populations of cells. Xenotransplantation assays functionally validated LSC-enrichment in $GPR56^{+}$ and $NKG2DL^{-}$ populations via engraftment in immunocompromised mice. Of note, engraftment was almost exclusively restricted to $GPR56^{+}$ and $NKG2DL^{-}$ populations. These different sorted populations from all SyTASC samples were submitted to multi-omics profiling (RNA-seq and DNA methylation) (Figure 6).

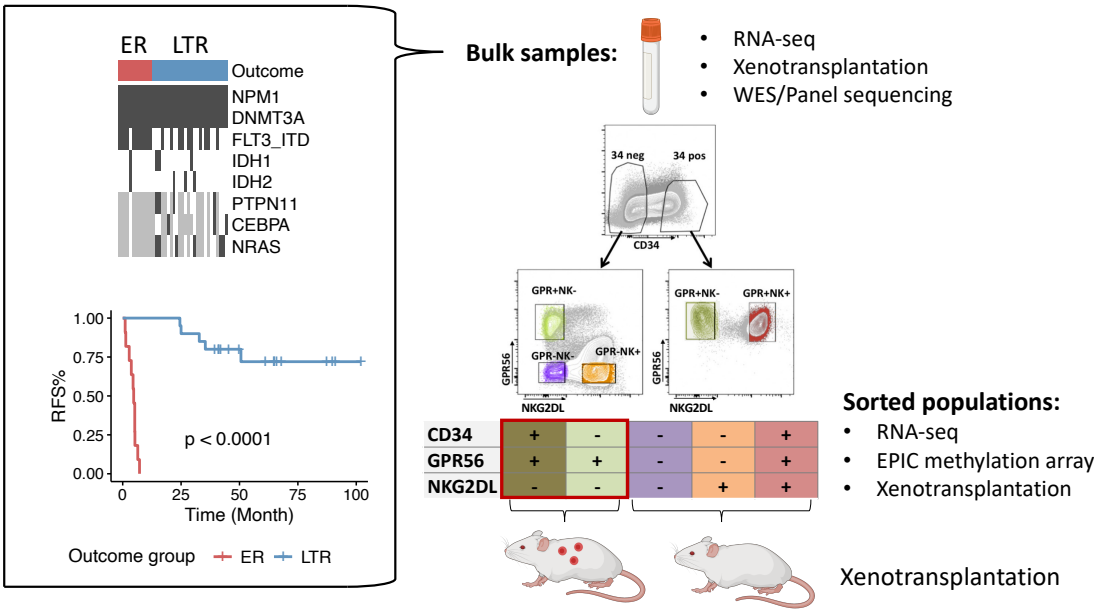

Figure 6: Simplified overview of the SyTASC project, the FACS sorting strategy, and sample profiling. The left panel depicts the genetic homogeneity of the SyTASC cohort. The heatmap at the top shows the most frequently mutated genes stratified by outcome group. Dark grey: mutated, white: wild type, and light gray: NA. The Kaplan-Meier curve at the bottom displays the different relapse-free survival (RFS) for the outcome groups. The "bulk samples" were profiled by RNA-seq, xenotransplantation assays, and mutational analysis by panel or whole exome sequencing (WES). The right panel displays a representative example of the sorting strategy including markers CD34, GPR56, and NKG2DL. The "sorted populations" were profiled by RNA-seq, DNA methylation, and xenotransplantation assays. The engrafting LSC-enriched populations are colored in green tones and highlighted by a red box. The FACS plot was created by Dr. Nadia Correia.

1.7 Aims of this thesis

The SyTASC cohort and the novel sorting strategy target two main challenges of AML research. (i) Enrichment of LSCs is crucial to identify relevant biological mechanisms for effective treatment. LSCs produce all leukemic cells and often present resistance to standard therapy. Therefore, their complete therapeutic eradiation is essential for a lasting cure. The novel marker combination successfully enriches for LSCs as functionally validated by xenotransplantation assays. (ii) The prognosis of AML patients is strongly associated with leukemic driver mutations and further factors such as age and cytogenetic aberrations. Hence, patient stratification is essential to identify suitable therapies. The SyTASC cohort is balanced and genetically homogeneous (*DNMT3A*-R882 and *NPM1*-W288fs*12), allowing to study biological processes in an unconfounded data set and to identify relapse mechanisms. Despite the homogeneous background, there is a remaining unexplained part of the variance in outcome as some patients presented with early relapse while others achieved long-term remission. Identification of relapse mechanisms is key to identifying better treatment options for these patients.

This unique cohort was submitted to multi-omics profiling (RNA-seq, DNA methylation, and genetic information). The aim of this thesis is the computational investigation of this data set to identify biological processes that pre-dispose for early relapse by primarily focusing on LSC-enriched cell populations acquired by the novel sorting strategy. The objectives thus are:

(i) Characterize the five different populations derived from the novel FACS sorting strategy, including two engrafting LSC-enriched populations.

(ii) Investigate the differences between the outcome groups (ER and LTR), including the identification of biological processes facilitating early relapse that could potentially be used as therapeutic targets as well as a stratification of patients by training an outcome prediction signature.

An additional aspect of the work was the metabolic analysis of the LSC-enriched populations. (iii) While working with metabolic data, in a technical side project, an interactive platform was set up to infer suitable extraction protocols for metabolomics studies.

2 Results

The AML SyTASC data set provides different levels of information. First, the different data types and second, the different sorted populations of which two have already been characterized as LSC-enriched in xenotransplantation experiments. In the following sections, I present an unsupervised integrative analysis of the data set, characterize the sorted cell populations based on the various omics data layers and present the differences between the two outcome groups (ER: early relapse; LTR: long-term remission). Finally, I will present metabolomics data as another layer to investigate the differences between ER and LTR samples and, based on another smaller independent data set, a technical platform to infer suitable extraction protocols for metabolomics studies.

2.1 Major drivers of variability: unsupervised integration of omics data

Transcriptome, methylation, as well as mutation information were integrated using Multi-Omics Factor Analysis (MOFA) to infer major drivers of variability in the data set (Figure 7a). Even though the total variance explained across all latent factors (LTs) was slightly lower for the transcriptome, LT1 is mainly driven by transcriptomic differences (Figure 7b,c). This factor, representing the primary source of variance, is strongly associated with the different sorted populations (Figure 7d,e). As shown in Figure 7e, there is a gradient between NKG2DL- and NKG2DL+ populations when clustering genes driving the variability encoded in LF1. Besides the striking difference between the different cell types represented by LF1 and LF15, most other LFs were associated with multiple clinical and biological sample features. (Figure 7d). MOFA results also showed that the intra-patient variability is highly associated with most LFs (Figure 7d). Hence, this confounding factor was taken into consideration in subsequent analysis whenever possible. The outcome group (ER vs. LTR) was clearly associated with LF3, LF10, and LF12. LF3 also showed a significant association with *FLT3*-IDT-mutant samples, whereas this effect was minor in LF10 and LF12 (Figure 7d, also cf. Figure 58). Taken together, the unsupervised multi-omics analysis showed that cell population and outcome group, as the two main dimensions of interest, were major drivers of variability in the data set. The next two sections aim to elucidate the biology of these effects in more detail.

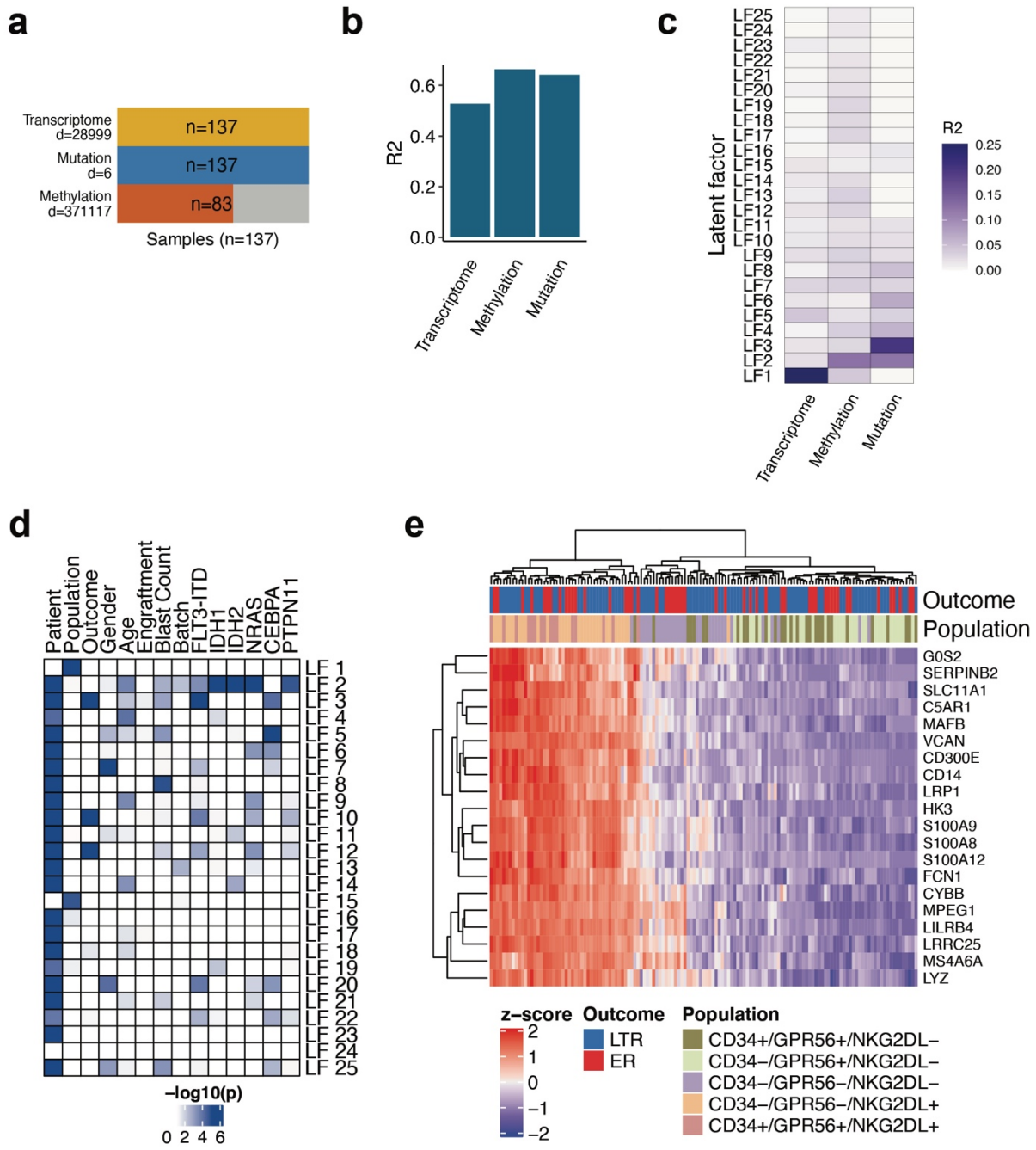

Figure 7: Integration of multi-omics data with MOFA. a) Data layers, number of features, and coverage across all samples (n=137). Of note, mutated genes were only included if present in at least three samples (cf. Figure 58) b) Total variance explained per data layer. c) Variance explained per latent factor. d) Associations of LFs with clinical and biological sample features. Color coding according to to p-values of Kruskal-Wallis Rank Sum Tests between respective groups and LFs. Engraftment: leukemic engraftment of bulk samples. e) Expression of genes with highest loadings for the transcriptomic layer in LF1.

2.2 A novel sorting strategy: enrichment for LSCs

The novel sorting strategy to enrich for LSCs, implemented by Dr. Nadia Correia, was functionally validated by xenotransplantation experiments in NOD scid gamma (NSG) mice by Dr. Elisa Donato. LSCs were enriched in $CD34^{+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$ populations, while $CD34^{-}GPR56^{-}NKG2DL^{-}$, $CD34^{-}GPR56^{-}NKG2DL^{+}$ and $CD34^{+}GPR56^{+}NKG2DL^{+}$ populations did not show engraftment in these experiments (Table S 1). Transcriptomic analysis of marker genes used for flow cytometry showed that immunophenotypes were also largely reflected by gene expression. However, some NKG2D ligands presenting very low expression were not significantly altered between the sorted populations (Figure S 2).
The following section aims to characterize the five sorted populations based on transcription and methylation data. I could show that $GPR56^{+}NKG2DL^{-}$ LSCs are highly different from the other, more differentiated populations and present a phenotype strongly reflective of malignancy and stemness. While the $CD34^{+}$ LSC population appears to contain, in addition to LSCs, retained healthy and pre-leukemic HSCs, the $CD34^{-}$ LSC population showed exclusive enrichment of LSCs.

2.2.1 Sorted populations are highly different based on transcription and methylation

Global analysis of RNA-seq and 850k array methylation data by principal component analysis (PCA) showed strong differences between $NKG2DL^{-}$ and $NKG2DL^{+}$ populations. Both data layers showed that the distinct cell types in the sorted populations account for more variability than the difference between leukemic and healthy samples (Figure 8). Interestingly, for $NKG2DL^{-}$ populations, leukemic and healthy samples clearly separated along principal component (PC) 2 for RNA-seq whereas the $NKG2DL^{+}$ populations were intermingled (Figure 8a).

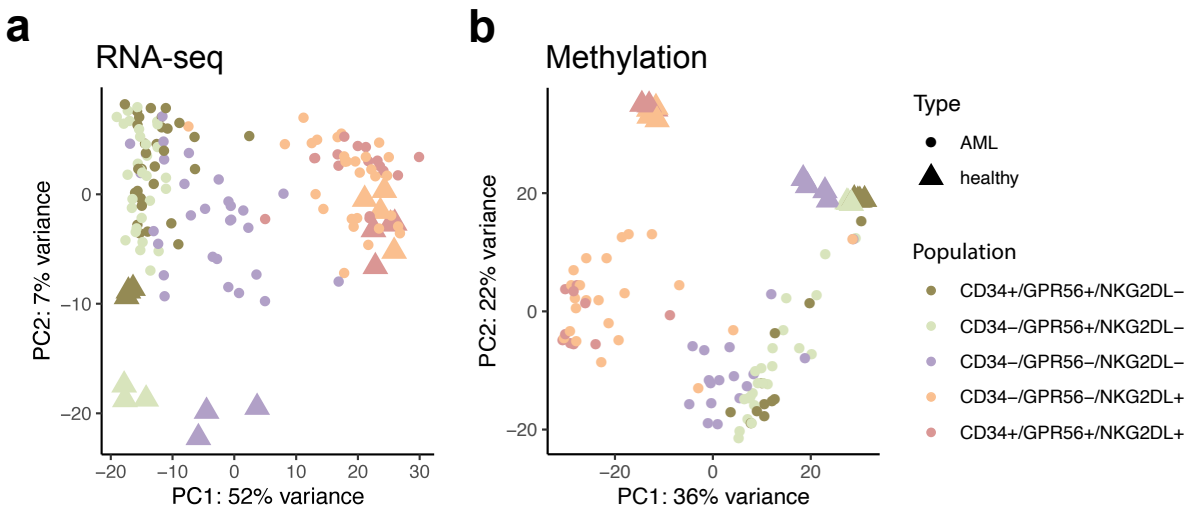

Figure 8: Variability between sorted populations of healthy compared to leukemic samples. a) PCA of RNA-seq. b) PCA of methylation data.

Consensus partitioning also showed stable separation between NKG2DL$^-$ and NKG2DL$^+$ populations in RNA-seq and methylation data (Figure 9a,b). This observation confirmed that the used markers enriched for very distinct cell populations. Clustering the most variable expression and beta values displayed the same pattern, respectively. Interestingly, for the methylation data, this seemed to be mainly driven by CpG positions in open sea regions (Figure 9c). Enrichment analysis of this CpG cluster, located in open sea (marked by a black frame), showed that these positions are significantly bound by regulatory genes involved in hematopoietic differentiation (Figure 9d). This unsupervised clustering already indicated that NKG2DL$^-$ populations were enriched for more immature populations based on methylation and displayed striking differences in clustering. However, only the GPR56$^+$NKG2DL$^-$ populations could be functionally validated to be enriched for LSCs as shown by the xenotransplantation assays.

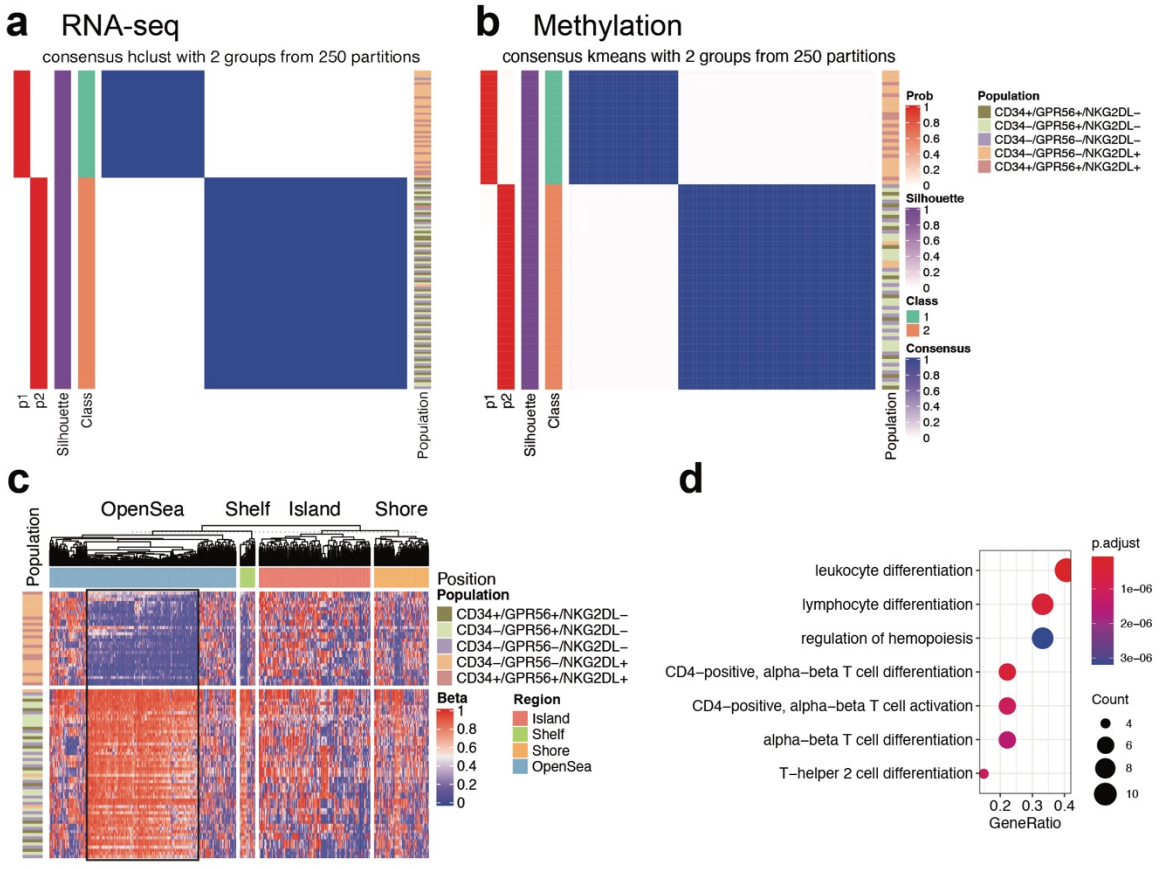

Figure 9: Consensus clustering shows a striking difference between sorted cell populations. a) Consensus clustering of RNA-seq data. b) Consensus clustering of methylation data. c) Clustering of the 1000 most variably methyleted positions. Positions in the cluster marked by the black frame were submitted to LOLA (Locus Overlap Analysis). d) Over-representation analysis of binding regulatory genes inferred by enrichment for genomic region sets estimated by LOLA.

2.2.2 LSC-enriched populations are more stem-like and cycle faster

HSPCs are morphologically different from more differentiated cells.[178] This is also reflected by the FACS measurements FSC A (forward scatter area) and SSC A (side scatter area). As observed for the clustering shown above, the $CD34^-GPR56^-NKG2DL^-$ population showed more similarity to the engrafting LSC-enriched population ($CD34^+GPR56^+NKG2DL^-$ and $CD34^-GPR56^+NKG2DL^-$) presenting a smaller diameter and a lower granularity (Figure 10).

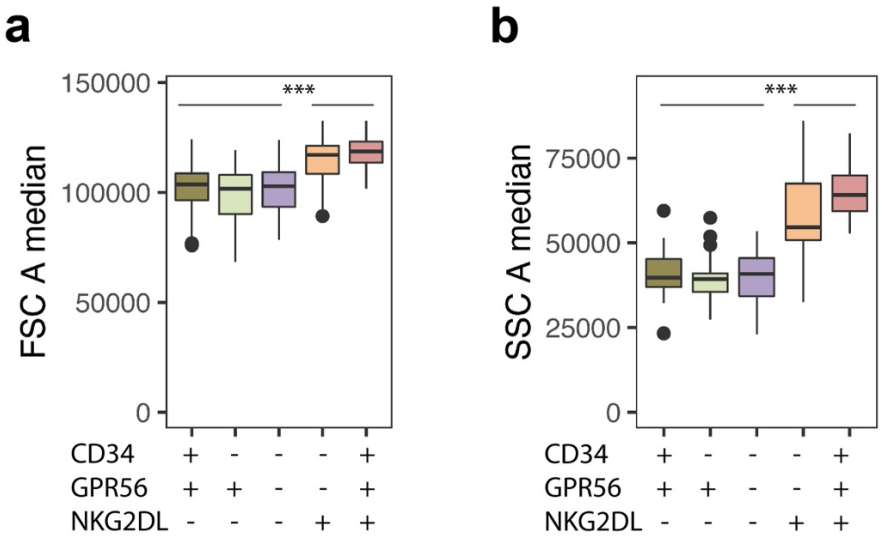

Figure 10: Morphology of sorted populations represented by the FACS measures FSC and SSC. a) FSC reflecting the diameter of cells. b) SSC reflecting the granularity of cells. Statistical differences were calculated between groups using a one-way ANOVA: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

However, when investigating LSC signatures published by Ng et al., the LSC17 signature scores and the correlation with the 104-gene LSC signature genes were significantly higher in the engrafting LSC-enriched compared to the non-engrafting populations. (Figure 11a,b).[51] As previously observed with PCA, populations appear to exhibit an increasing gradient of differentiation from the stem-like populations $CD34^+GPR56^+NKG2DL^-$ and $CD34^-GPR56^+NKG2DL^-$ over the populations $CD34^-GPR56^-NKG2DL^-$, $CD34^-GPR56^-NKG2DL^+$ to the most differentiated population $CD34^+GPR56^+NKG2DL^+$ (Figure 11a,b and cf. Figure 8). The observation that engrafting populations are more stem-like was also confirmed by gene set enrichment analysis (GSEA). Comparing engrafting and non-engrafting populations showed significant enrichment of the Eppert LSC signature.[50] On the contrary, the gene set for "KEGG Hematopoietic Lineage" was enriched in non-engrafting populations indicating that these cells are more differentiated (Figure 11c).

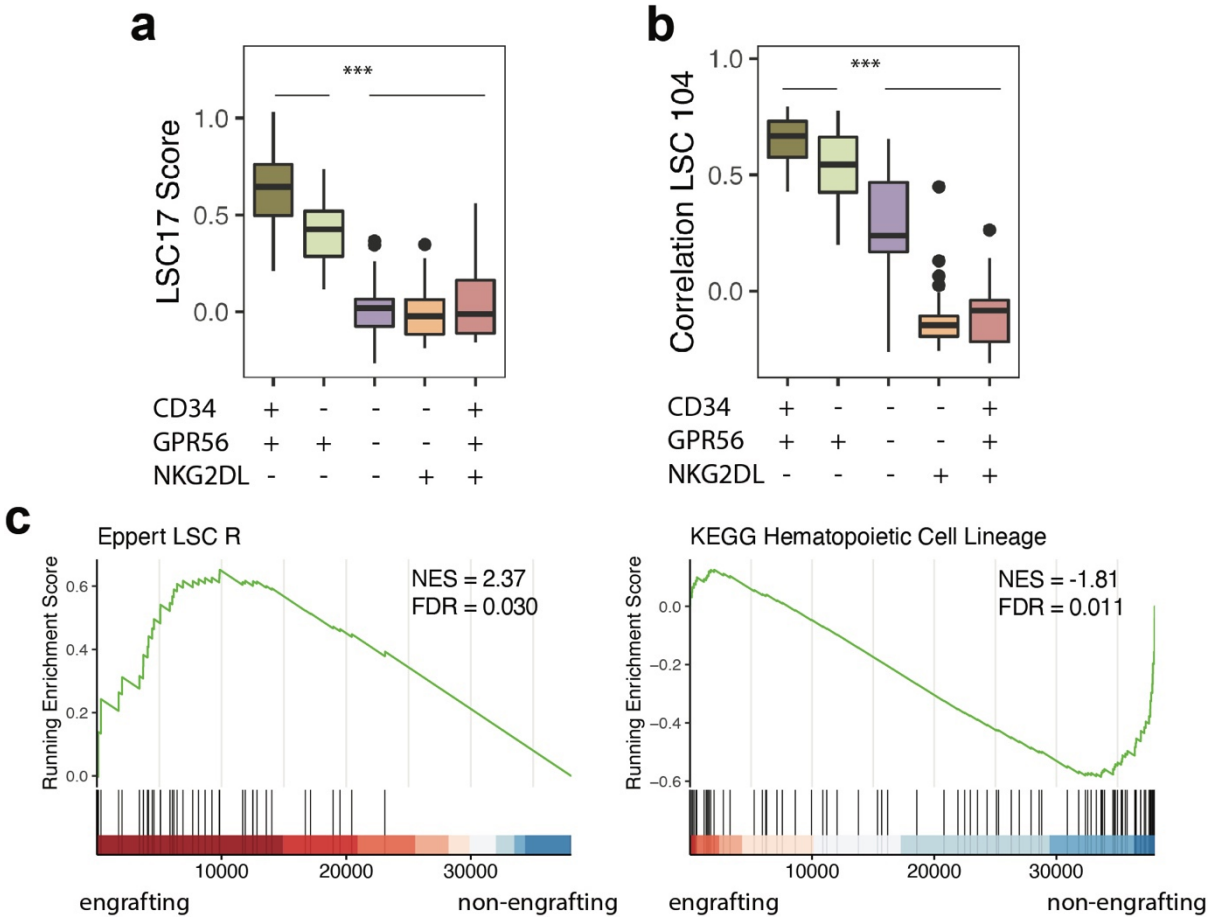

Figure 11: LSC signatures and gene sets are specific for engrafting LSC-enriched populations. a) LSC17 signature score. b) Correlation with 104-genes LSC signature. Statistical differences were calculated between groups using one-way ANOVA: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant. c) GSEA plots.

The hypothesis, that engrafting LSC-enriched populations are more stem-like was further supported by embedding the sorted populations into single-cell data of healthy human hematopoiesis.[40] The LSC populations CD34$^+$GPR56$^+$NKG2DL$^-$ and CD34$^-$GPR56$^+$NKG2DL$^-$ clustered with healthy HSCs and progenitor cells, in accordance with their stem-like phenotype. The CD34$^-$GPR56$^-$NKG2DL$^-$ population did not show a clear pattern, whereas CD34$^-$GPR56$^-$NKG2DL$^+$ cells clustered to promonocytes or monocytes, and the most differentiated CD34$^+$GPR56$^+$NKG2DL$^+$ population clustered clearly to monocytes (Figure 12). When embedding the sorted populations derived from healthy bone marrow samples, the clustering pattern was similar for the differentiated populations. However, CD34$^+$GPR56$^+$NKG2DL$^-$ and CD34$^-$GPR56$^+$NKG2DL$^-$ did not cluster to HSCs but to erythroid progenitor cells. This may indicate that the novel sorting strategy is only valid for leukemic samples, even though the more differentiated populations clustered similarly for leukemic

and healthy samples (Figure S 3). Interestingly, some of the CD34^{-}GPR56^{+}NKG2DL^{-} and CD34^{-}GPR56^{-}NKG2DL^{-} samples clustered to NK cells (Figure 12).

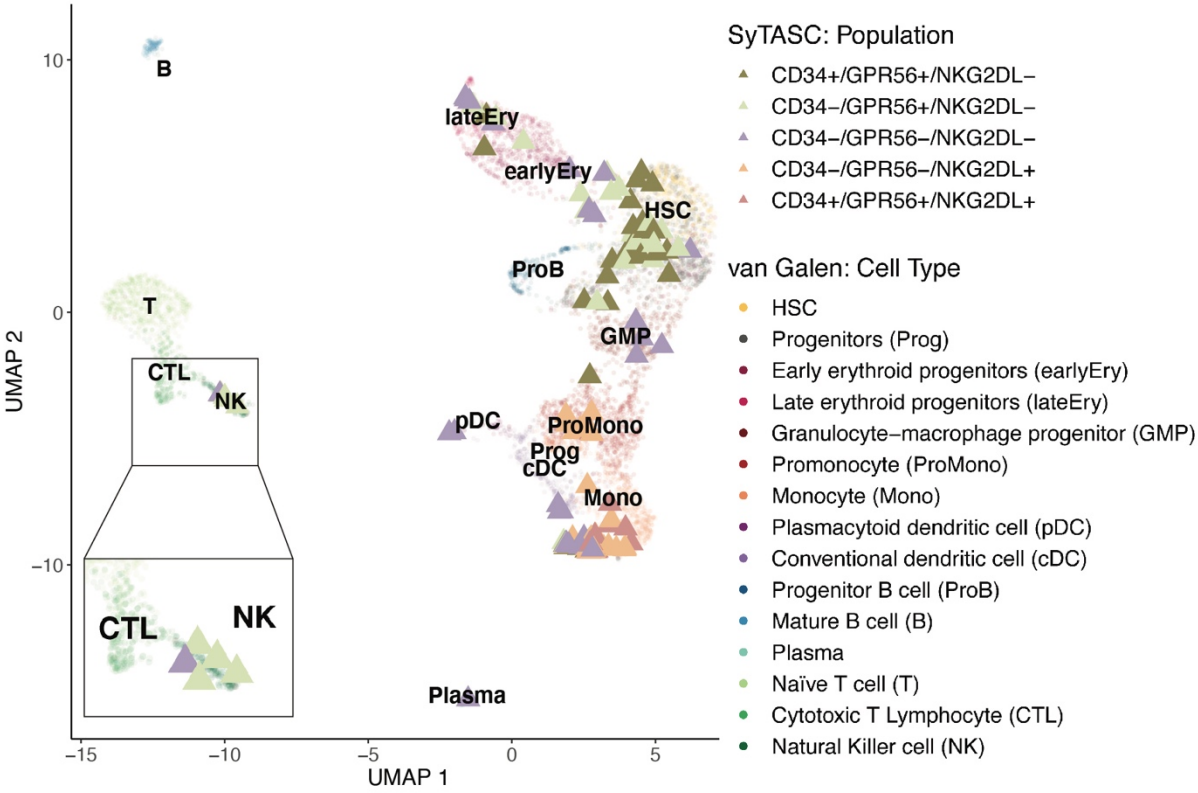

Figure 12: Embedding of sorted populations into UMAP of single-cell data published by van Galen et al. [40]**.** Box in the left bottom shows a representative magnification of CTL and NK populations.

Since the SyTASC cohort was selected for *DNMT3A* and *NPM1* double-mutated samples, I investigated the expression and mutant allele frequency of these genes. Both genes, *DNMT3A* and *NPM1*, showed significantly higher counts in LSC-enriched populations than in the more differentiated cells (Figure 13b,d). Contrarily, the mutant allele frequency of *NPM1* was significantly increased in the more differentiated samples while for *DNMT3A* no difference in allele frequency was observed (Figure 13a,c). Overall, the mutant allele frequencies were on average higher for *DNMT3A* (around 50%) compared to *NPM1* (below 40%). Consequently, the higher frequency of *DNMT3A* indicated a higher abundance of *DNMT3A*-mutant clones, while *NPM1* is not present in all cells. Notably, *DNMT3A* mutant allele frequency showed a large variability.

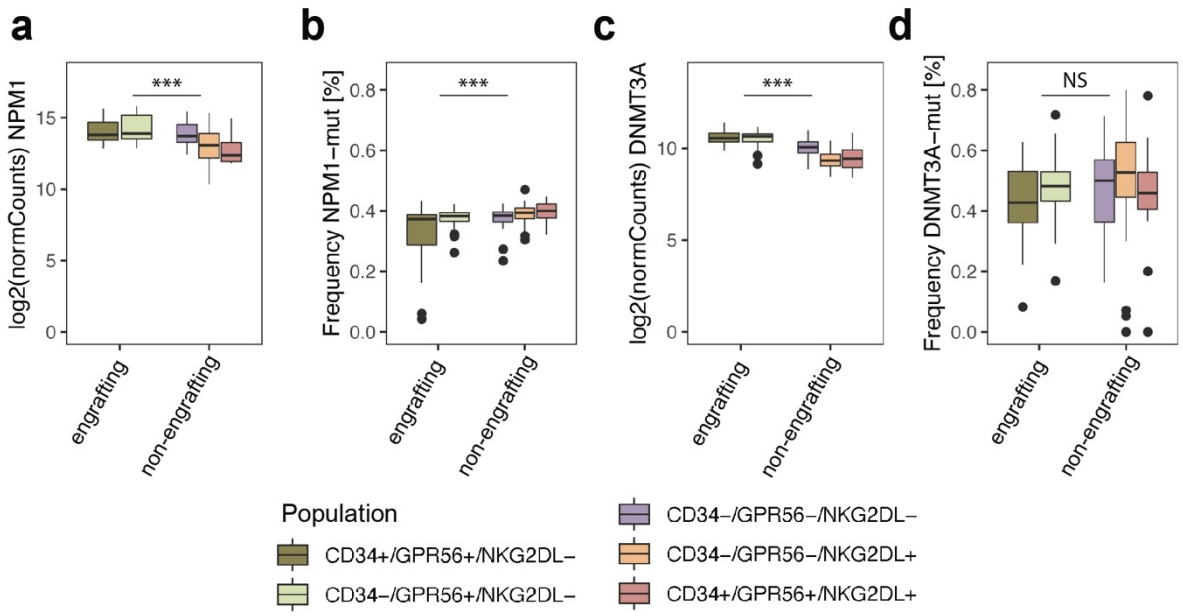

Figure 13: Box plots showing mutant allele frequency and expression of *NPM1* and *DNMT3A*. Data were stratified by engrafting LSC-enriched populations compared to more differentiated populations. a) Mutant allele frequency of *NPM1* based on RNA-seq data. b) Expression of *NPM1*. c) Mutant allele frequency of *DNMT3A* based on RNA-seq data. b) Expression of *DNMT3A*. Statistical differences were calculated between groups using a one-way ANOVA: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

When investigating the global methylation pattern of the sorted population, I observed a similar differentiation gradient as described above. Across all regions and positions, increasing differentiation was associated with decreasing methylation (Figure 14a). This effect was mainly driven by open sea, shelf, and shore regions. Since island regions were almost completely demethylated, there was no difference between the sorted populations (Figure 14b-e). Additionally, a data set published by Jung et al. was used to compare the sorted populations to probe-specific methylations patterns of healthy HSCPs. In contrast to RNA-seq data which showed clear differences between the sorted populations, all sorted population from AML samples shared the highest similarity with methylation pattern of GMPs (Figure S 4 and cf. Figure 12).

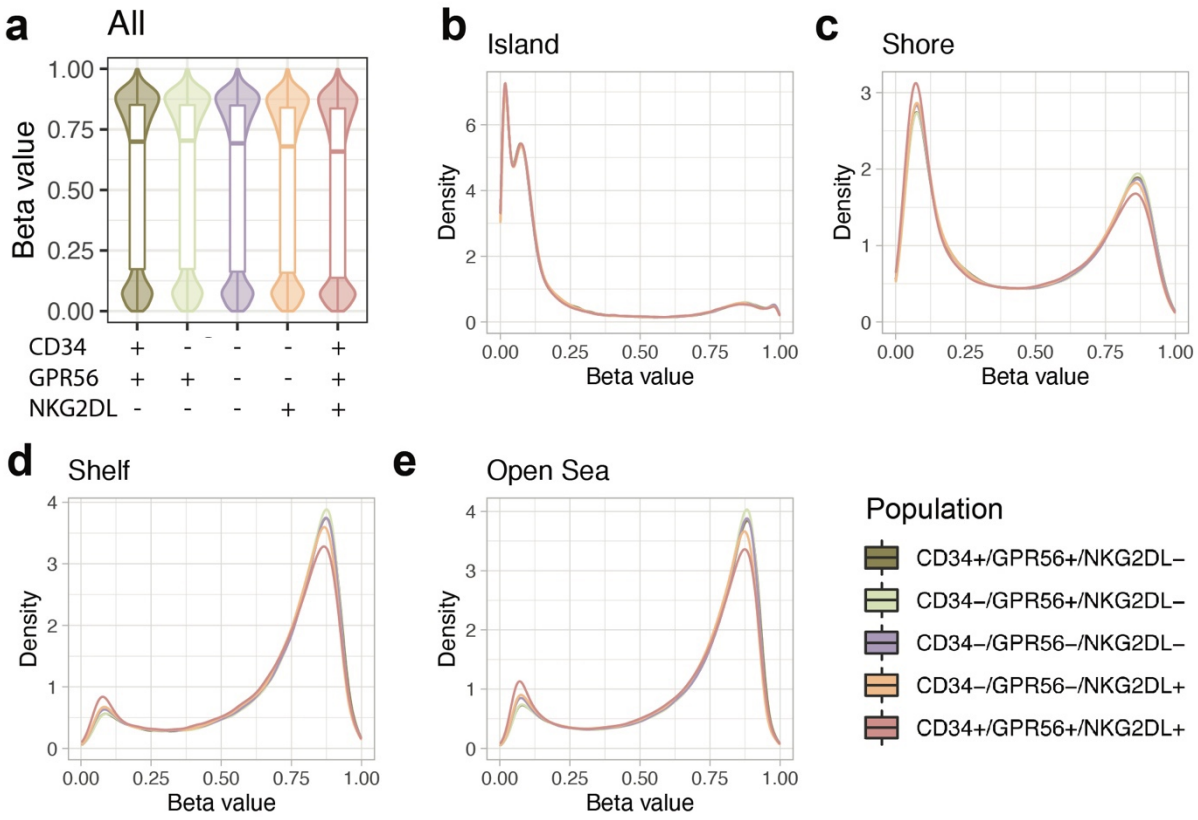

Figure 14: Global methylation across sorted populations. Distribution of beta values across a) all regions, b) island regions, c) shore regions, d) shelf regions, and e) open sea regions. Color code for populations applies to all items.

Inferring cell cycle phases revealed that LSC samples were significantly more often in S phase than more differentiated samples (Figure 15a). Whereas the G2M score was not significantly different between the sorted populations, the S score was significantly higher in LSC populations. Again a gradient from LSC towards the more differentiated populations could be observed (Figure 15b,c). Accordingly, GSEA showed significant enrichment for cell cycle-related processes in LSC populations, exemplarily represented by the curated gene sets “Fisher G1 S Cell Cycle”, “Whitfield Cell Cycle G1 S”, “Cell Cycle DNA Replication”, and “WP G1 to S Cell Cycle Control” (Figure 16a). In line with this observation, gene sets related to DNA replication and strand elongation showed significant enrichment in LSC-enriched populations indicating a higher replication rate (Figure 17a).

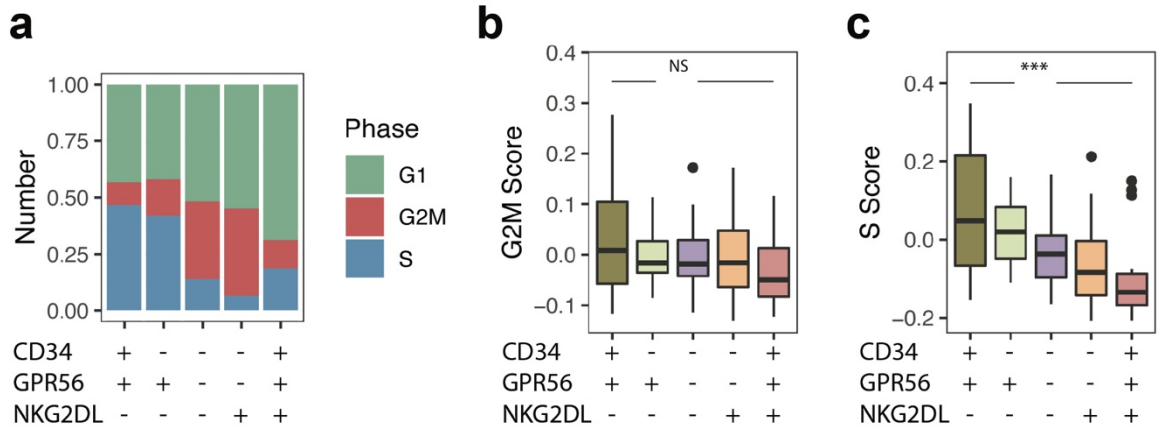

Figure 15: Estimation of cell cycle phases. a) Stacked bar plot of inferred cell cycle phases based on G2M and S score. Fisher's Exact Test on numbers of samples: p-value = $1.5x10^{-3}$. Pairwise Fisher's Exact Test between engrafting and non-engrafting population: p-value = $4.2x10^{-5}$. b) G2M score. c) S score. Statistical differences were calculated between groups using a one-way ANOVA: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

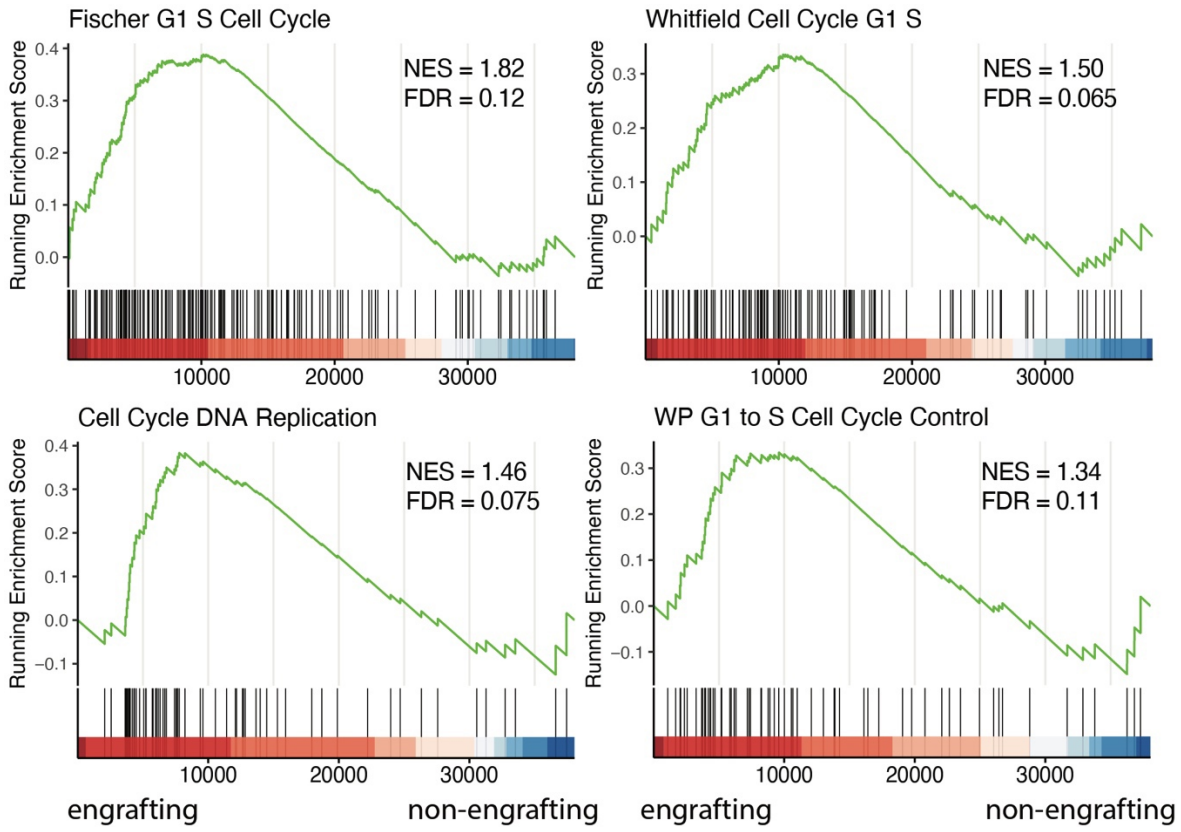

Figure 16: GSEA for cell cycle-related gene sets. GSEA plots for "Fisher G1 S Cell Cycle", "Whitfield Cell Cycle G1 S", "Cell Cycle DNA Replication", and "WP G1 to S Cell Cycle Control".

2.2.3 Expression of immunoregulatory genes and pathways in differentiated populations

The direct comparison of the populations revealed marked enrichment of gene sets related to inflammatory processes in more differentiated, non-engrafting compared to engrafting LSC-enriched populations (Figure 17a). Enrichment of the most significantly different gene sets was strongly driven by Leukocyte Immunoglobulin-Like Receptors (e.g., *LILRB1*) but also S100 calcium-binding proteins (e.g., *S100A9*) and different interleukins (Figure 17b). The expression was strongly associated with the sorted population and significantly different between engrafting and non-engrafting cells (Figure 17c , Figure S 5, and also cf. LF1 of MOFA analysis in Figure 7e).

Notably, also immunoglobin genes were differentially expressed between the sorted populations. Several immunoglobulins were markedly expressed in the LSC populations, while others showed high expression in the $CD34^{-}GPR56^{-}NKG2DL^{-}$ population (Figure S 6). Immunoglobin transcripts have been described to originate from ambient mRNA and therefore might not represent the actual expression of the respective population.[179] The expression of immunoglobins appeared in multiple leading-edge analyses when performing GSEA in this study. Given the possibility that these expression differences may be erroneous, I further investigated the origin of the mRNA by comparing the frequency of spliced and unspliced transcripts. Gaidatzis et al. described a computational approach for this analysis, by quantifying intronic and exonic reads.[180] The frequency of intronic counts was almost neglectable and originated mainly from one sample indicating that the expression of immunoglobulins were rather ambient mRNA than of intracellular origin (Figure S 7). Consequently, immunoglobulins were removed from all subsequent analyses in this study.

Investigating the expression of stimulatory and inhibitory immune checkpoints showed population-specific expression for some genes (Figure 17d). The stimulatory checkpoints *TNFRSF4* (also known as *CD134* or *OX40* receptor) and *TNFRSF18* (also known as glucocorticoid-induced TNFR-related protein (*GITR*) or *CD357*) were highly expressed in $NKG2DL^{-}$ populations. Contrarily, the inhibitory checkpoints *CYBB*, *HAVCR2*, *C10orf54*, and *SIGLEC7* were highly expressed in $NKG2DL^{+}$ populations. Additionally, multiple Killer Cell Immunoglobulin Like Receptors (KIRs) were specifically expressed in the $CD34^{-}GPR56^{+}NKG2DL^{-}$ population. This was in line with the observation of some $CD34^{-}GPR56^{+}NKG2DL^{-}$ samples clustering to NK cells when embedded in healthy hematopoiesis (cf. Figure 12).

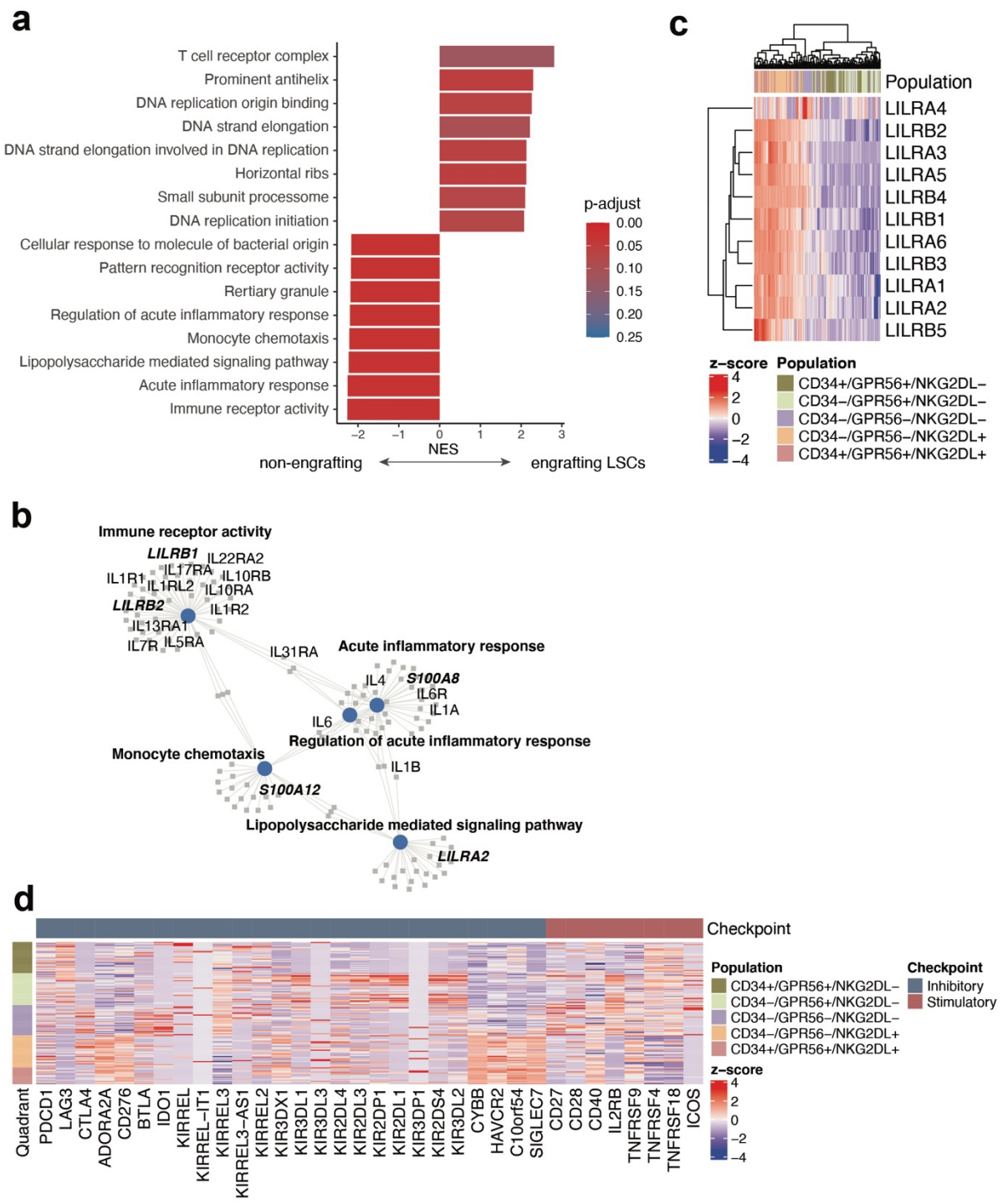

Figure 17: Immunoregulatory processes are highly enriched in differentiated populations. a) Bar plot showing GSEA gene sets with highest normalized enrichment scores (NES) between engrafting and non-engrafting populations. b) Selected genes identified by leading-edge analysis for the five most enriched gene sets. c) Heatmap showing expression of LILRs. d) Heatmap showing expression of inhibitory and stimulatory immune checkpoints.

2.2.4 Engrafting populations are transcriptionally very similar

The two LSC-enriched populations were further investigated to identify transcriptional differences. However, as described above, various analyses showed high abundancy of NK cells in the $CD34^{-}GPR56^{+}NKG2DL^{-}$ population, and also differential expression analysis of these populations was corrupted by the presence of these cells (cf. Figure 12, Figure 17e, and Figure S 8). Therefore, Dr. Elisa Donato re-sorted populations from the SyTASC cohort using an improved sorting strategy that depleted lineage-positive cells including NK cells. This approach aimed at both, increasing the purity of LSCs and allowing comparison between populations. The percentage of NK cells in both data sets were inferred by cell type deconvolution of bulk tissue using the data set published by Corces et al. as a reference.[20] The proportion of NK cells in the improved "NK-depleted" data set was marginal compared to the "original" data set described above (Figure 18). Hence, the improved data set was suitable to compare the two engrafting populations on a transcriptional level.

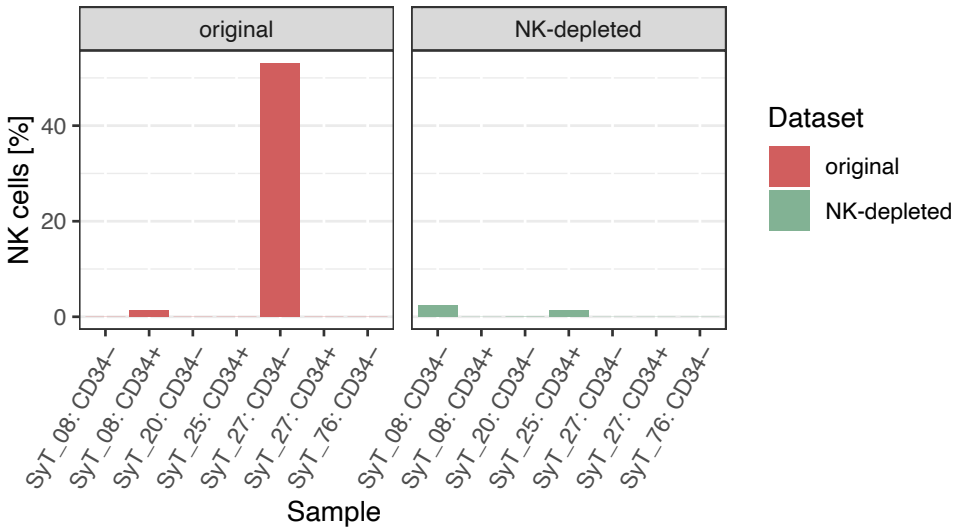

Figure 18: Proportion of NK cells in selected samples used for re-sorting by improved FACS sorting strategy. Left: "original" samples as used for analyses above. Right: "NK-depleted" samples sorted by improved FACS protocol.

An initial PCA of engrafting AML LSC-enriched populations and their healthy counterparts showed that the AML populations clustered closely together, whereas healthy populations clustered separately far apart along PC1 (Figure 19a). As described above, embedding into healthy hematopoiesis indicated that the sorting strategy mainly enriched for erythroid progenitor cells when applied to healthy samples (Figure S 3). However, the LSC17 score and the correlation with the 104-genes LSC signature were comparably high for the healthy

CD34$^+$GPR56$^+$NKG2DL$^-$ population which also clustered closer to the AML LSC-enriched populations (Figure 19a-c). GSEA on PC1 showed enrichment of gene sets related to HSCs, LSCs, and *NPM1*-mutated AML for the two AML and the healthy CD34$^+$GPR56$^+$NKG2DL$^-$ populations whereas healthy CD34$^-$GPR56$^+$NKG2DL$^-$ showed enrichment for G2 cell cycle-related gene sets (Figure S 9). Hence, the healthy CD34$^+$GPR56$^+$NKG2DL$^-$ population seemed to be enriched for HSCs; likely because of the positive selection for CD34. This was also supported by the enrichment for HSC-related gene sets (e.g., Jaatinen hematopoietic stem cell up) in the healthy CD34$^+$GPR56$^+$NKG2DL$^-$ population compared to the AML samples (Figure 19d). Interestingly, normalized enrichment scores (NES) were very similar for both LSC-enriched populations (CD34$^+$GPR56$^+$NKG2DL$^-$ and CD34$^-$ GPR56$^+$NKG2DL$^-$) when compared to the healthy HSC-enriched (CD34$^+$GPR56$^+$NKG2DL$^-$) population. However, especially the LSC17 score was significantly lower in the CD34$^-$GPR56$^+$NKG2DL$^-$ LSC-enriched AML population (Figure 19b).

As indicated by the proximity in the clustering, the similarity of the LSC-enriched populations was also reflected in a direct comparison by a low number of differentially expressed genes (99). But when deliberately investigating differential aspects, a high activity of the SRY-Box Transcription Factor 2 (*SOX2*) in the CD34$^+$GPR56$^+$NKG2DL$^-$ population was observed, which is essential for self-renewal and embryonic stem cell maintenance (Figure 20a,b).[181] Consistently, estimation of key pathway activity also showed striking enrichment for JAK-STAT signaling in the CD34$^+$ population, which is also related to the regulation of hematopoiesis and HSC proliferation (Figure 20c,d).[165] Taken together, the significantly higher LSC17 score, comparable to healthy HSCs, in the CD34$^+$GPR56$^+$NKG2DL$^-$ AML population and the differential pathway activity led to the hypothesis this population might contain healthy and pre-leukemic HSCs.

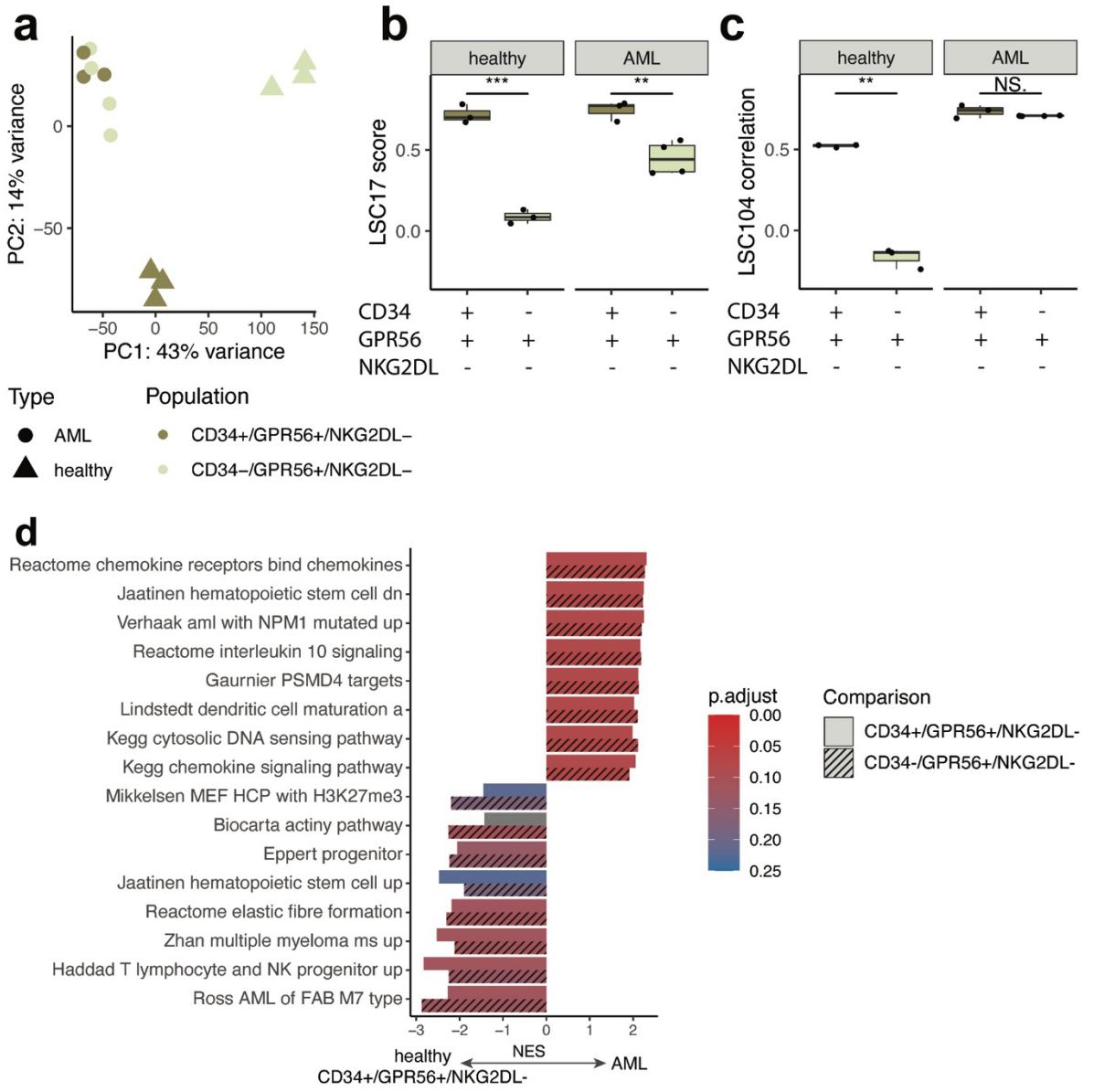

Figure 19: Engrafting LSC populations are similar when compared to healthy counterparts. a) PCA of healthy and AML $CD34^{+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$ populations. b) LSC17 score. c) Correlation with LSC 104 genes. Statistical differences were calculated between groups using a two-sided Student's t-Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant. d) GSEA between healthy $CD34^{+}GPR56^{+}NKG2DL^{-}$ population and AML $CD34^{+}GPR56^{+}NKG2DL^{-}$ or $CD34^{-}GPR56^{+}NKG2DL^{-}$ populations.

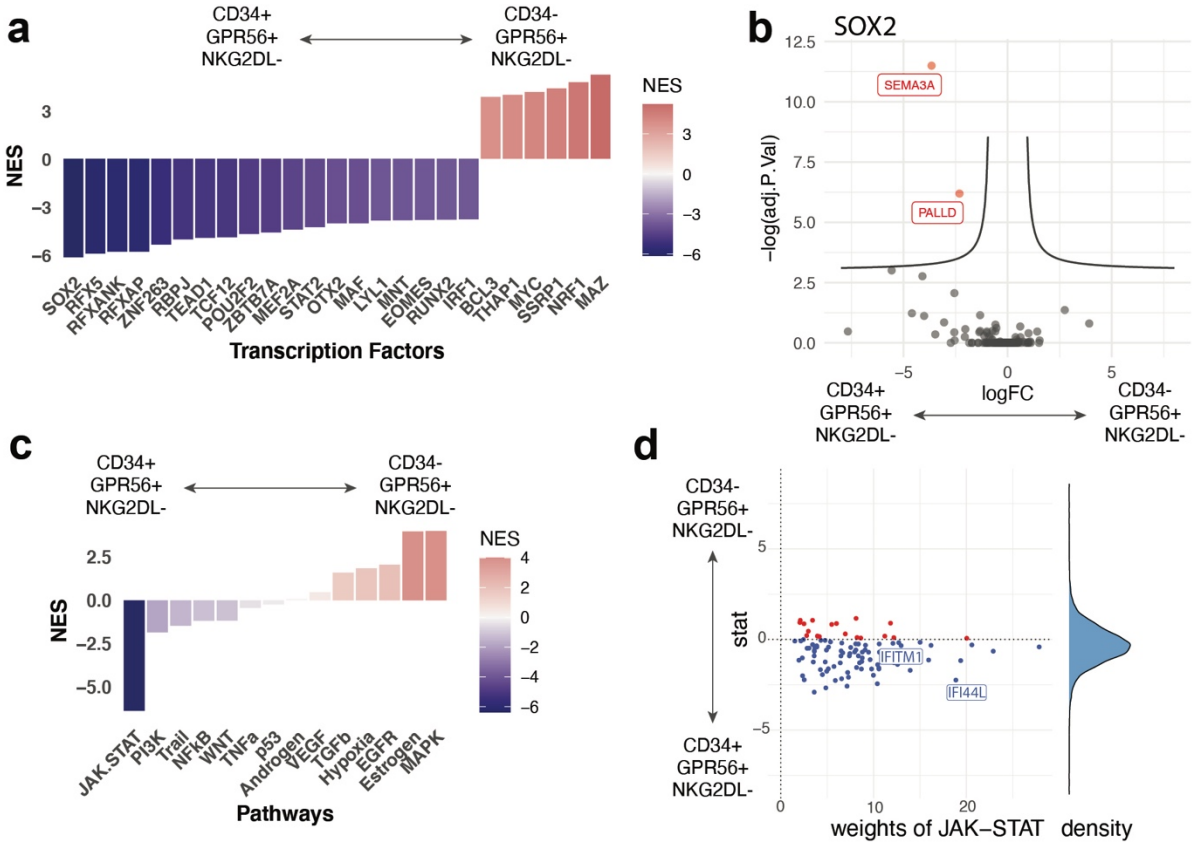

Figure 20: Activity of transcription factors and key pathways. Activity was inferred by VIPER and PROGENy between CD34$^-$GPR56$^-$NKG2DL$^-$ and CD34$^+$GPR56$^+$NKG2DL$^-$ populations. a) NES for 25 most differentially active transcription factors. b) Exemplary volcano plot for genes driving the enrichment for transcriptions factor SOX2. c) NES for key pathways. d) Exemplary expression and weights of genes driving the activity of the JAK-STAT pathway.

2.2.5 CD34$^+$ LSC populations contain healthy retained and pre-leukemic HSCs

To further investigate the hypothesis of retained HSCs, the abundance of healthy HSCs in the samples was inferred using deconvolution with a data set of healthy hematopoietic cells published by Corces et al. as a reference.[20] The fraction of healthy HSCs was clearly higher in the CD34$^+$GPR56$^+$NKG2DL$^-$ population compared to the CD34$^-$GPR56$^+$NKG2DL$^-$ population from AML samples. However, due to the low number of samples in the "NK-depleted" data set, statistical significance was not achieved. The deconvolution approach was also supported by fractions calculated in healthy samples. As expected, the proportion of HSC was highest in the healthy CD34$^+$GPR56$^+$NKG2DL$^-$ samples, while healthy HSCs were absent in the healthy CD34$^-$GPR56$^+$NKG2DL$^-$ population (Figure 21a). Consistently, HSC signatures were significantly enriched in CD34$^+$GPR56$^+$NKG2DL$^-$ samples compared to CD34$^-$GPR56$^+$NKG2DL$^-$ AML samples (Figure 21b). In line with the hypothesis, some of the CD34$^+$GPR56$^+$NKG2DL$^-$ samples

presented multi-lineage engraftment in xenotransplantation experiments, whereas CD34$^-$GPR56$^+$NKG2DL$^-$ exclusively initiated in AML engraftment (Figure 21c, cf. Table S 1).

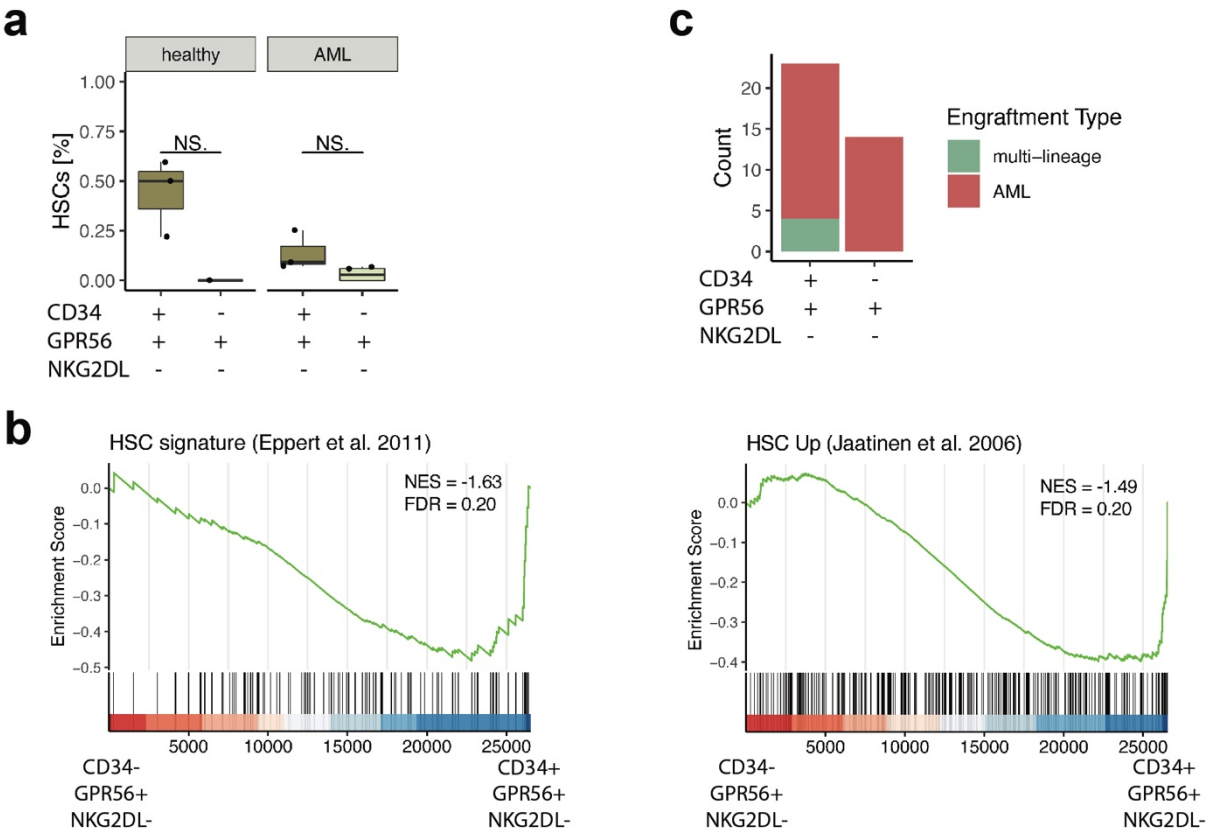

Figure 21: Retained healthy HSCs in CD34$^+$GPR56$^+$NKG2DL$^-$ population. a) Percentage of healthy HSCs inferred by deconvolution using Corces et al. as a reference [20]. Statistical differences were calculated between groups using a two-sided Student's t-Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant. b) Selected GSEA plots showing enrichment for the HSC signatures in the CD34$^+$GPR56$^+$NKG2DL$^-$ compared to CD34$^-$GPR56$^+$NKG2DL$^-$ AML populations.[50,182] c) Stacked bar plot showing engraftment type of LSC populations in the "original" data set.

Mutant allele frequencies are a proxy for the abundance of leukemic, pre-leukemic and healthy cells, and therefore mutant allele frequencies were assessed in the different cell populations. In the CD34$^+$GPR56$^+$NKG2DL$^-$ population, the mutant allele frequency of *NPM1* was lower compared to CD34$^-$GPR56$^+$NKG2DL$^-$ cells. Again, statistical significance was lacking in the "NK-depleted" data set. Still, the "original" data set showed a significantly lower percentage (Figure 22a,c). The mutant allele frequency of *DNMT3A* was not statistically different and only showed a minor trend in the "original" data set (Figure 22b,d). Interestingly, samples with very low mutant allele frequencies in the "original" data set showed rather multi-lineage engraftment in xenotransplantation experiments. However, assessment of the

engraftment type has not been performed for all samples (Figure 22c,d and cf. Figure 21c). This observation further supported the hypothesis of retained healthy and pre-leukemic HSCs in the CD34$^+$GPR56$^+$NKG2DL$^-$ AML population.

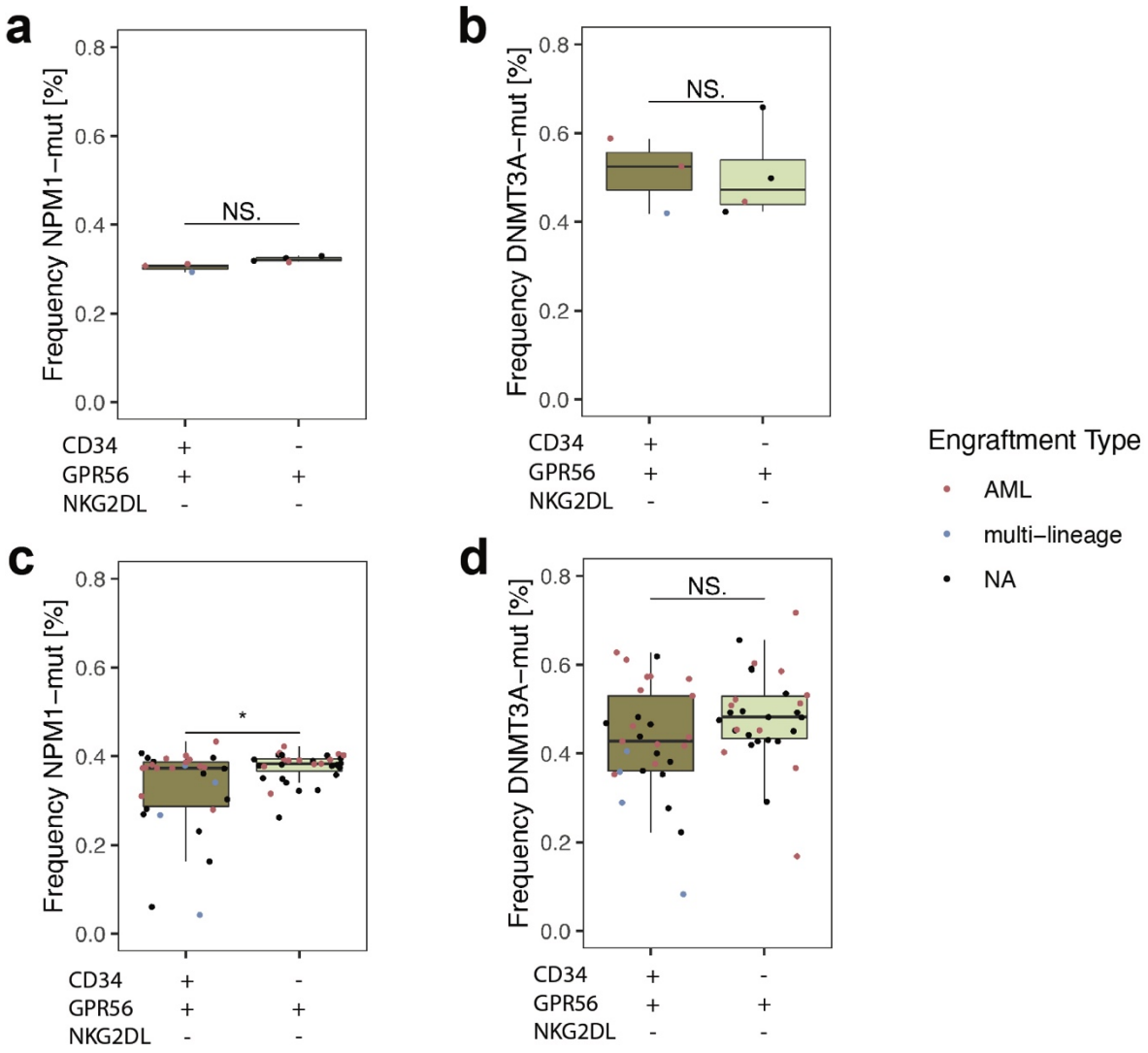

Figure 22: Mutant allele frequency in RNA-seq data from CD34$^+$GPR56$^+$NKG2DL$^-$ and CD34$^-$GPR56$^+$NKG2DL$^-$ AML populations. a) Percentage of *NPM1*-mutated counts in the "NK-depleted" cohort. b) Percentage of *DNMT3A*-mutated counts in the "NK-depleted" cohort. c) Percentage of *NPM1*-mutated counts in the "original" cohort. d) Percentage of *DNMT3A*-mutated counts in the "original" cohort. Statistical differences were calculated between groups using a two-sided Student's t-Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

2.2.6 Difference between CD34$^-$GPR56$^-$NKG2DL$^-$ and engrafting LSC-enriched populations

Several analyses in the previous sections indicated similar phenotypical properties of the CD34$^-$GPR56$^-$NKG2DL$^-$ and the engrafting LSC-enriched populations (CD34$^+$GPR56$^+$NKG2DL$^-$

and CD34$^-$GPR56$^+$NKG2DL$^-$). For example, these cells were morphologically very similar based on size and granularity. In Addition, consensus clustering revealed two very distinct clusters separated by NLG2DL status (less differentiated NKG2DL$^-$ populations and more differentiated NKG2DL$^+$ cells; cf. Figure 9 and Figure 10). For most analyses the CD34$^-$GPR56$^-$NKG2DL$^-$ population was located at the center of the gradient between the stem-like LSC populations and the more differentiated ones. This was observed in PCA plots, in the analyses of cell cycle and methylation, in the embedding into healthy hematopoiesis as well as in immune-related expression (cf. Figure 8, Figure 12, Figure 14, Figure 15, and Figure 17). Also, when only the sorted AML populations were clustered, the described gradient was clearly represented in PC1 (Figure 23). However, according to the calculated LSC17 score, CD34$^-$GPR56$^-$NKG2DL$^-$ cells presented a less stem-like phenotype (Figure 11).

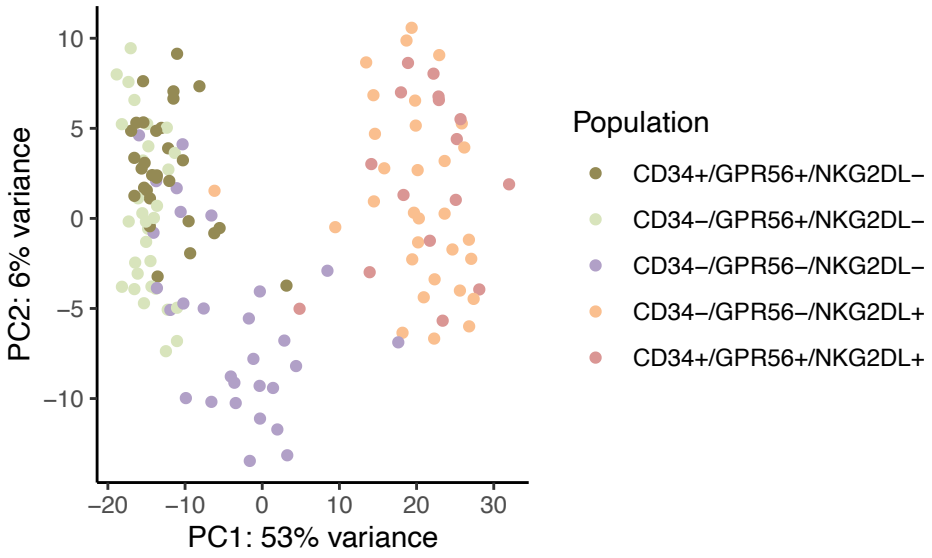

Figure 23: PCA of sorted populations from AML samples.

Since CD34$^-$GPR56$^-$NKG2DL$^-$ cells did not engraft in xenotransplantation experiments, a direct comparison between these cells and the engrafting populations was performed to identify processes that might explain the loss of engrafting potential (Table S 1). Analysis of transcription factor activity revealed enrichment for several genes involved in cell cycle and differentiation (Figure 24a). For example, Basic Helix-Loop-Helix Family Member E40 (*BHLHE40*), involved in differentiation, and E2F Transcription Factor 4 (*E2F4*), a suppressor of cell cycle activity, were enriched in the CD34$^-$GPR56$^-$NKG2DL$^-$ population (Figure 24b,c).[183–185] Conversely, Spi-1 Proto-Oncogene (*SPI1*) activity was enriched in the LSC populations leading to the downregulation of target genes such as BCL6 Transcription Repressor (*BCL6*) and Toll

Like Receptor 4 (*TLR4*), which are expressed in differentiated hematopoietic cells (Figure 24c).[186–188] This was also in line with GSEA results which showed enrichment for gene sets related to LSC, HSC, and G1 S cell cycle in the engrafting LSC populations (Figure 25). Thus, the loss of immunophenotypical GPR56 positivity, together with a loss of engraftment potential and LSC properties, was associated with lower stemness and lower cell cycle activity.

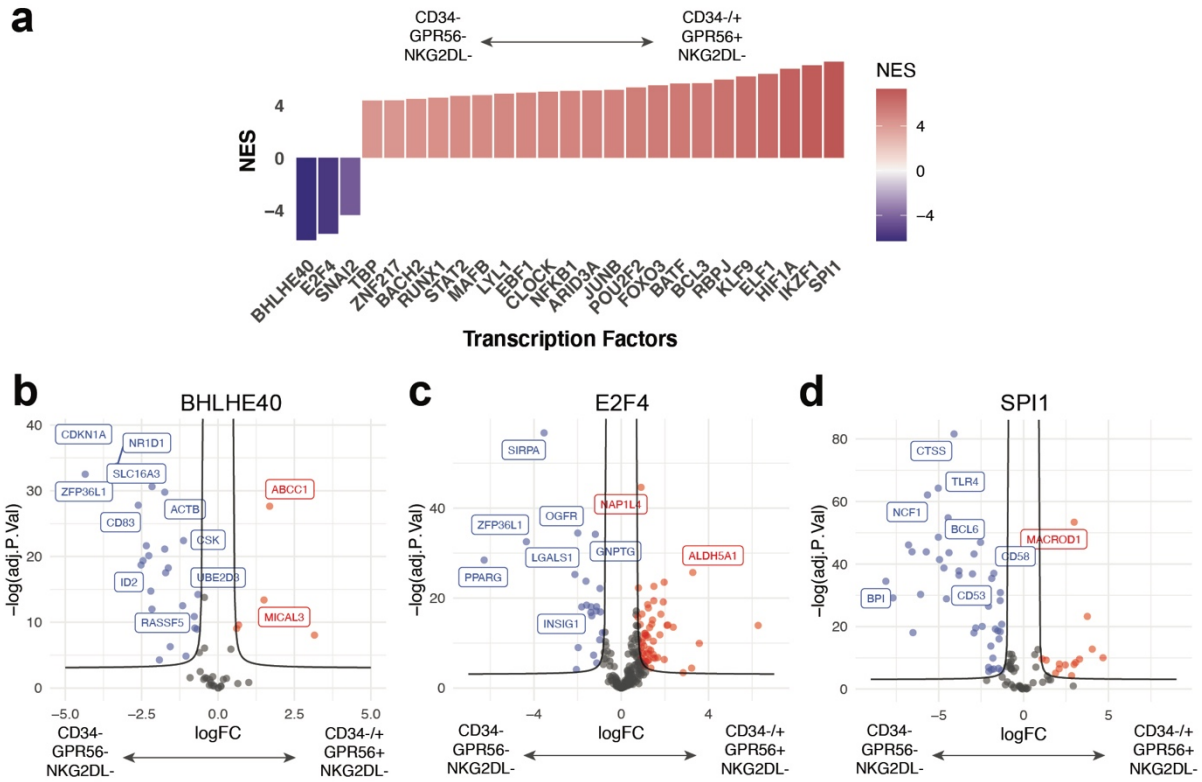

Figure 24: Activity of transcription factors inferred by VIPER between CD34$^-$GPR56$^-$NKG2DL$^-$ and engrafting LSC populations (CD34$^+$GPR56$^+$NKG2DL$^-$ and CD34$^-$GPR56$^+$NKG2DL$^-$). a) NES for 25 most differentially active transcription factors. Exemplary volcano plots for genes driving the enrichment for transcription factors b) BHLHE40, c) E2F4, and d) SPI1.

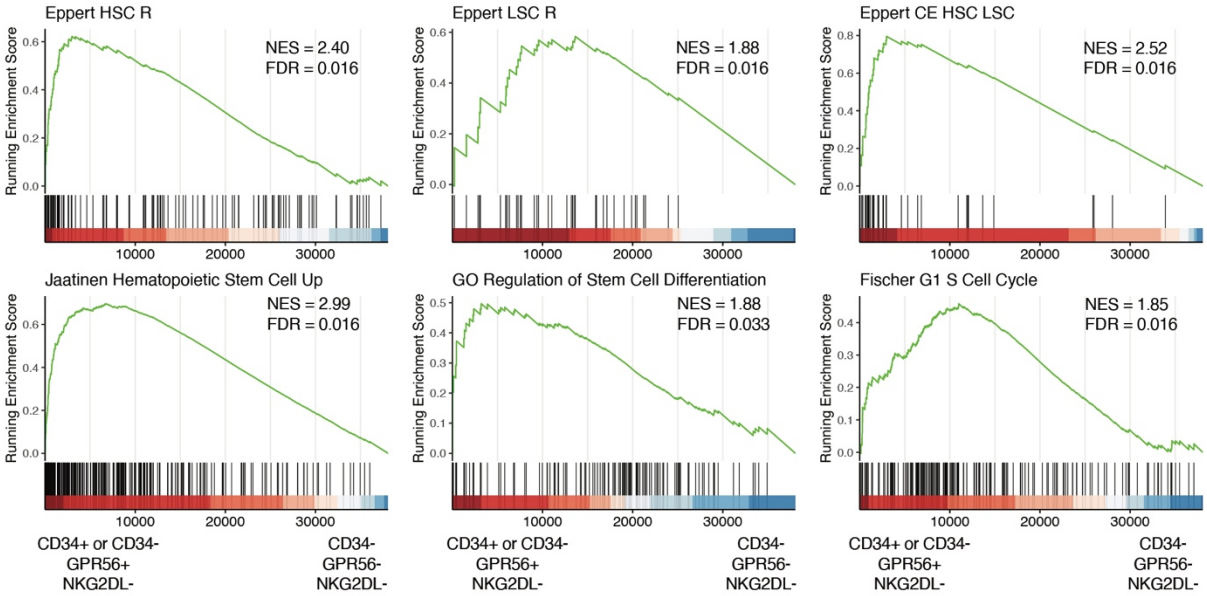

Figure 25: Enrichment of HSC, LSC, and cell cycle-related gene sets in engrafting populations. GSEA plots for selected gene sets between $CD34^{-}GPR56^{-}NKG2DL^{-}$ and engrafting LSC populations ($CD34^{+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$).

2.3 Two distinct outcome groups in a genetically homogenous cohort

After characterization of the different sorted populations, this section will focus on differences between the two outcome groups. Even though all patients went into complete remission after chemotherapy, 12 patients suffered from an early relapse (ER) while 26 achieved long-term remission (LTR). The first part will focus on potential confounding factors and show that phenotypical differences based on RNA-seq and methylation data were more pronounced in the engrafting LSC-enriched populations. These biological differences included changes in key pathways and energy metabolism which were confirmed in metabolomics analyses. A second part aimed to train an outcome prediction signature that exhibited high predictive power in multiple AML cohorts.

2.3.1 The SyTASC cohort has a homogeneous genetic background

The SyTASC cohort was retrospectively selected for *NPM1*- and *DNMT3A*-mutant patients. However, also other genes were mutated in subsets of the patients. To investigate whether those mutations were statistically associated with relapse-free survival (RFS), cox proportional hazard regression was used. Table 3 shows results of these analyses for frequently mutated genes. Only *FLT3*-ITD was significantly associated with RFS. Most, but not all ER patients were mutant for this gene, while only some LTR patients were *FLT3*-ITD-mutant (cf. Figure 58). Hence, this mutation seemed to be correlated with the RFS but could not fully explain the differential relapse.

Table 3: Statistical overview of mutated genes and their association with RFS. Genes were included if mutated in at least three samples or wild type in at least three samples. Hence, *NPM1* and *DNMT3A* were not included since these genes are mutated in all samples. Statistical parameters were calculated using a cox proportional hazards regression model. Mutations were not available for all samples indicated by the number of "NA". mut: mutant: wt: wild type.

Gene	Data	Beta	HR (95% CI for HR)	Wald test statistic	P-value
***FLT3*-ITD**	mut: 17, wt: 14	1.4	4 (1.3-12)	5.6	0.018
IDH1	mut: 4, wt: 27	0.31	1.4 (0.39-4.8)	0.23	0.63
IDH2	mut: 4, wt: 27	-0.14	0.87 (0.2-3.8)	0.04	0.85
PTPN11	mut: 4, wt: 9, NA: 18	-0.37	0.69 (0.071-6.6)	0.1	0.75
CEBPA	mut: 3, wt: 8, NA: 20	-21	1.1e-09 (0-Inf)	0	1
NRAS	mut: 5, wt: 8, NA: 18	0.03	1 (0.17-6.3)	0	0.97

2.3.2 Potential confounding factors do not explain the distinct outcome groups

Analogously to the mutational analysis, available clinical information was analyzed to identify potentially confounding factors. Age, gender, and white blood cells (WBC) were not significantly associated with the RFS. However, BM (bone marrow) blast counts were significantly correlated with relapse of patients. The abundance of BM blasts is a diagnostic parameter for AML and high levels have been described as an indicator of poor prognosis, likely due to an advanced progression.[189] Even though statistical significance was observed, the hazard ratio showed a low risk increase. Thus, the BM blast count may be considered to not be a major confounding factor.

Table 4: Statistics on available clinical information as potential confounding factors associated with RFS. Statistical parameters were calculated using a cox proportional hazards regression model. BM blast counts were not available for all samples indicated by the number of "NA".

Confounder	Data	Beta	HR (95% CI for HR)	Wald test statistic	P-value
Gender	m: 19, f: 12	0.016	1 (0.37-2.8)	0	0.98
Age	50 (22; 65)	0.024	1 (0.97-1.1)	0.88	0.35
WBC	59.8 (4.9; 261.5)	0.0019	1 (0.99-1)	0.2	0.65
BM blasts	75.75 (4; 93), 1 NA	0.031	1 (1-1.1)	4.8	0.028

2.3.3 Variability between outcome groups is more pronounced in LSC populations

Before focussing on the data set of sorted cell populations, analysis of bulk RNA-seq of whole patient samples was performed to identify transcriptional differences between the outcome groups. Gene expression analysis discovered that 40 genes were significantly differentially expressed (Figure 26), while GSEA did not reveal significant enrichment between the outcome groups (data not shown).

Similarly, when performing an analysis on a merged cohort consisting of all sorted populations from all patients, the outcome group was shown to be only a minor source of variability. Statistical association of PCA results with clinical information revealed that the outcome group was significantly linked to the variability represented by PC4 and PC8 (4% and 2% of the variability, respectively) (Figure 27a,b). A direct comparison between the outcome groups in merged cohorts consisting of all sorted populations showed that MHC-II-related gene sets

and antigen-presenting processes were enriched in ER samples, whereas processes related to cell adhesion were enriched in LTR samples (Figure 27c).

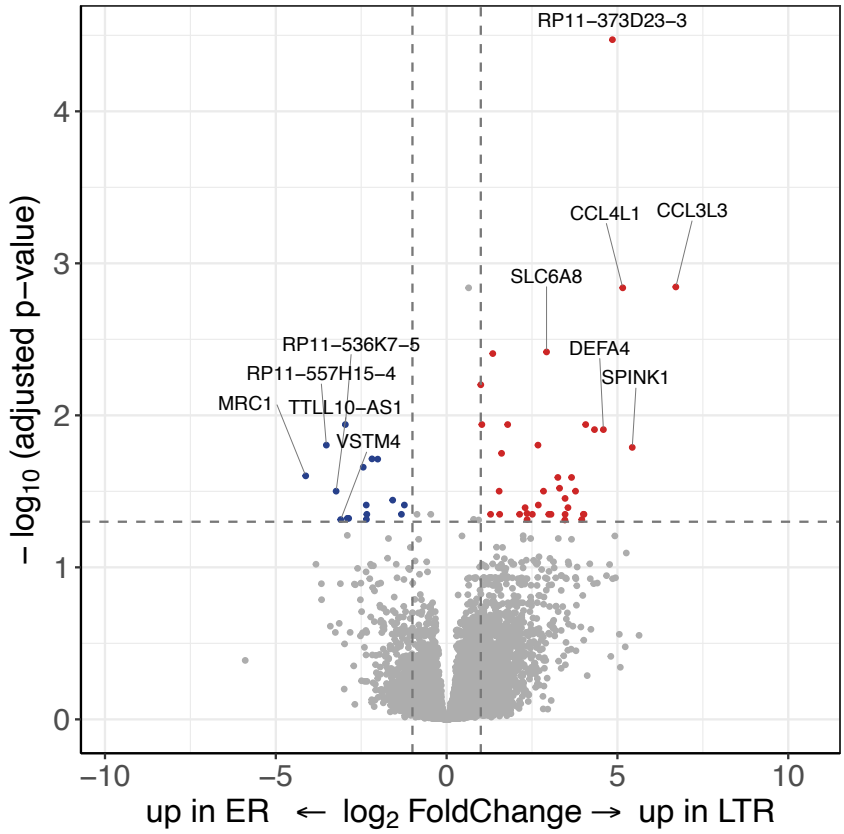

Figure 26: Volcano plot showing differentially expressed genes between ER and LTR samples in bulk RNA-seq.

Based on the definition and concept of leukemic stem cells, the LSC populations are most likely the one to be enriched for those cells re-initiating the disease after complete remission of patients. The cohort was therefore subset to these populations ($CD34^{+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$). Analogous analysis to investigate the sources of variability as for all populations revealed that PC1, PC3, and PC4 were associated with outcome and accounted for 15%, 9% and 7% of the variability, respectively (Figure 28). Notably, for PC1, the association with the outcome group was even stronger than the patient-specific effect. This observation was similar or even more pronounced in the analysis of the methylation data set (Figure S 10). Consequently, analyses presented in the following subsections focus on the engrafting LSC populations if not stated otherwise.

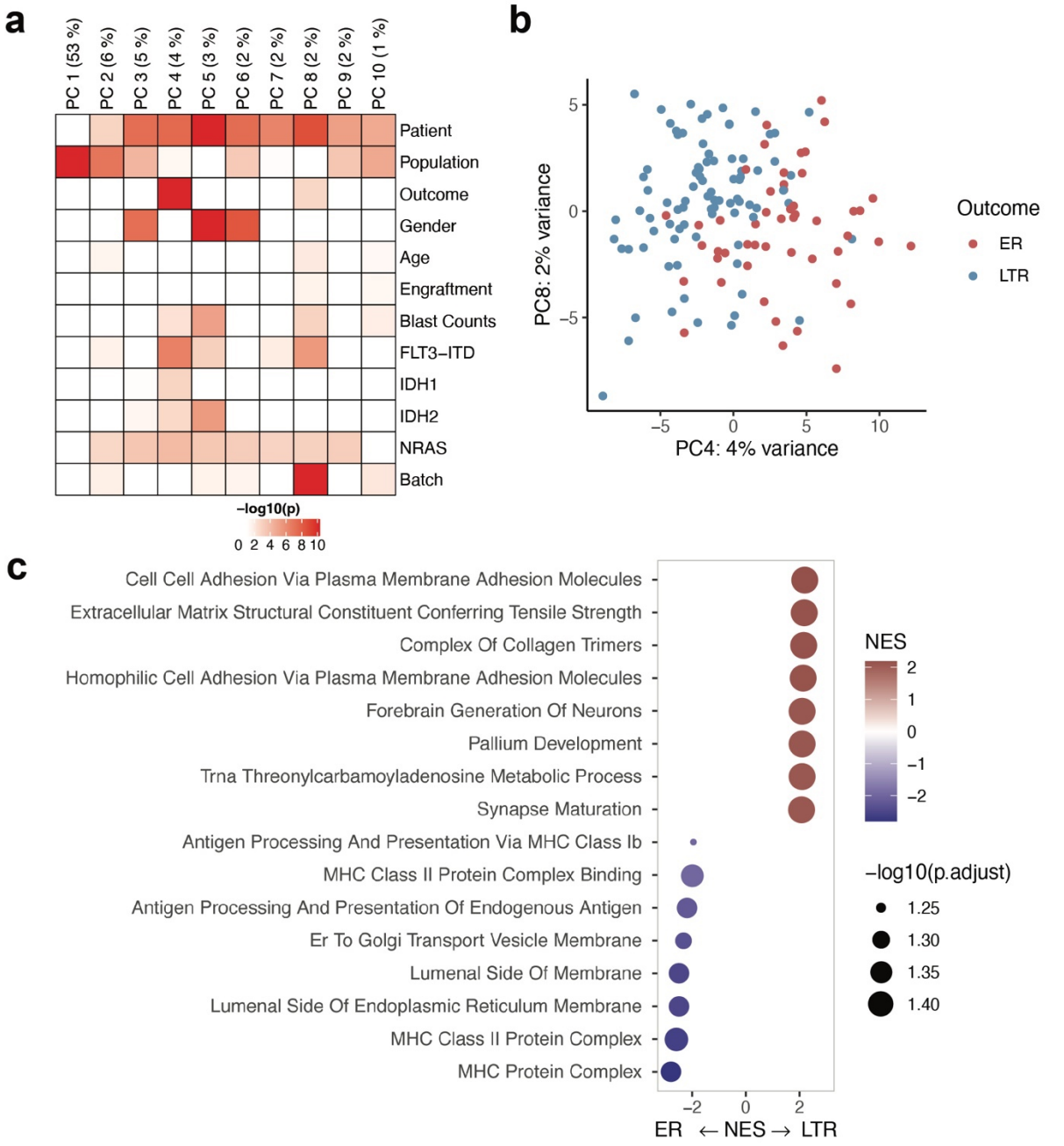

Figure 27: The outcome group is a minor source of variability in RNA-seq data across all sorted populations. a) Heatmap showing p-values of Kruskal-Wallis Rank Sum Test between biological or clinical information and first 10 principal components. b) Scatter plot of loadings for PC4 and PC8 (associated with the outcome groups). c) Bubble plot showing the most enriched gene sets among differentially expressed genes between outcome groups in a merged cohort consisting of all sorted populations from all patients.

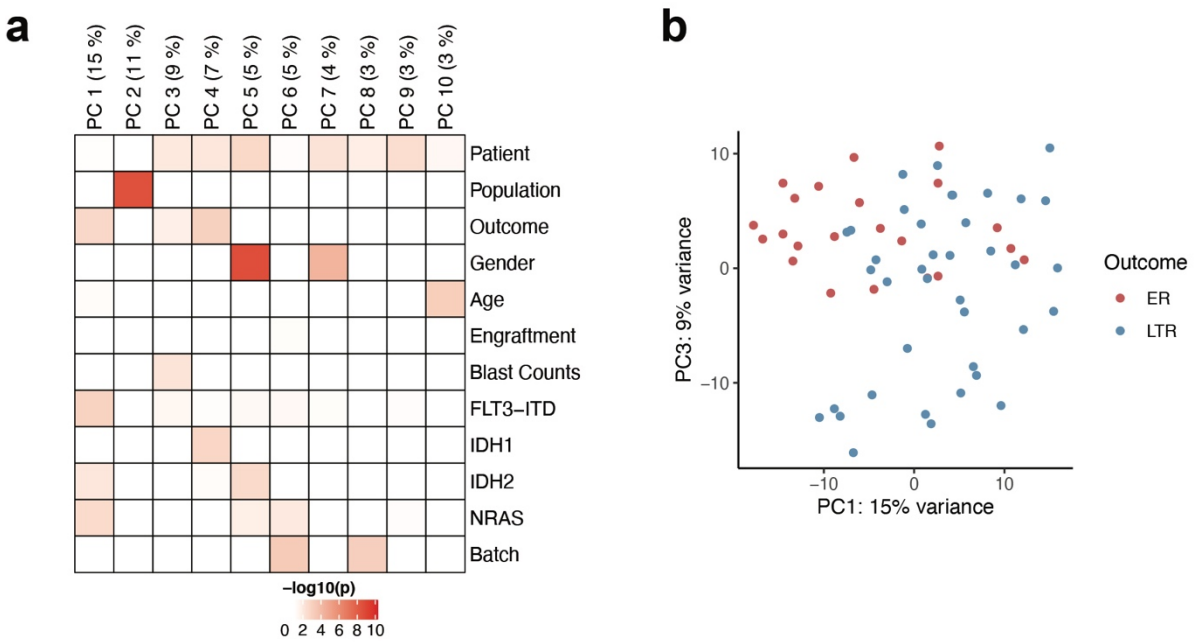

Figure 28: The outcome group is a major source of variability in RNA-seq data from engrafting LSC-enriched populations. a) Heatmap showing p-values of Kruskal-Wallis Rank Sum Test between biological or clinical information and first 10 principal components. b) Scatter plot of loadings for PC1 and PC3 associated with the outcome groups.

2.3.4 ER LSCs are more stem-like than LTR LSCs

A common hypothesis in cancer stem cell research is the quiescent or dormant phenotype of therapy-persistent cells allowing them to sustain chemotherapeutic intervention.[30] Therefore, potential differences in stemness between the outcome groups were investigated. The LSC17 score as well as correlation with the 104-gene signature were significantly higher in ER samples when analyzing the engrafting LSC populations. The LSC17 score calculated for the $CD34^{+}GPR56^{+}NKG2DL^{-}$ population showed a trend but was lacking significance. (Figure 29a,b). Consistently, gene sets related to LSCs and HSCs signatures were enriched in ER samples (Figure 29c). This observation was further supported by embedding the LSC-enriched populations into the healthy single-cell data set.[40] The relative median distance between healthy HSCs and ER samples (d = 1.62) was clearly lower than between HSCs and LTR samples (d = 2.75) (Figure 30).

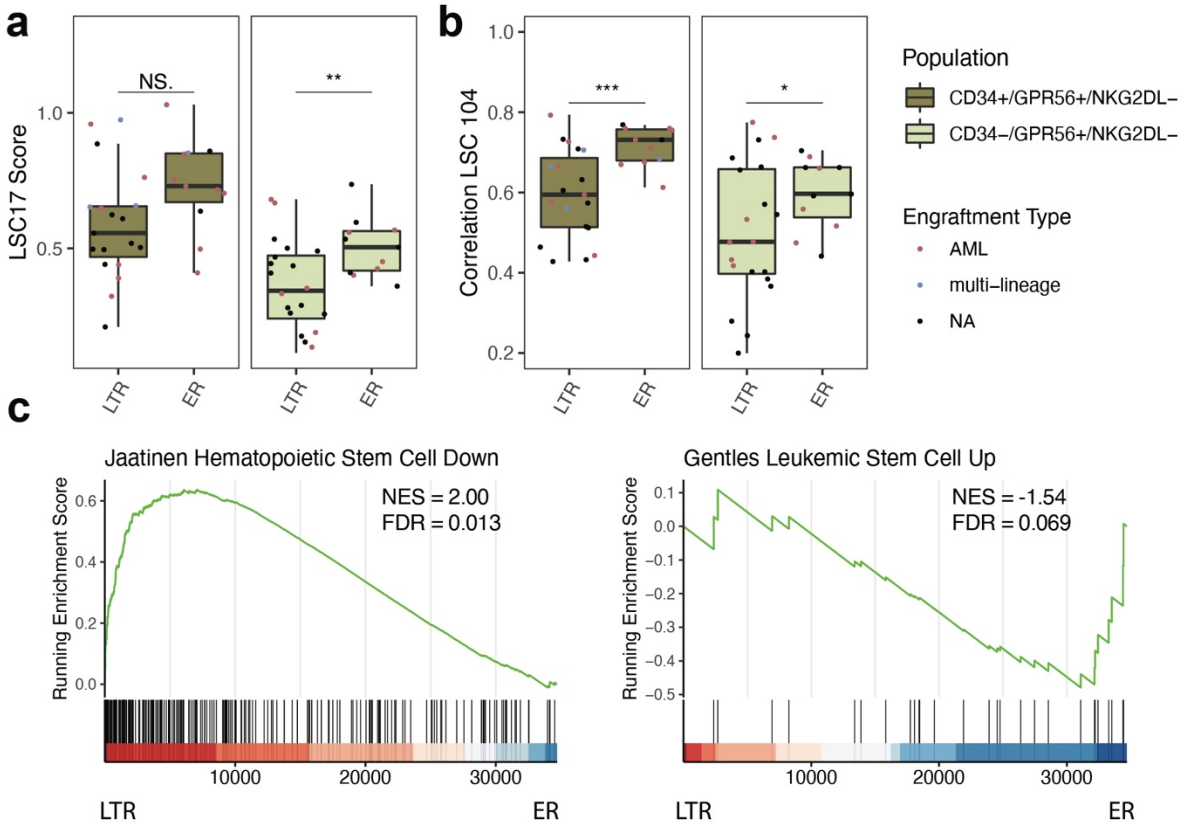

Figure 29: ER samples present a higher stem-like phenotype compared to LTR samples. a) LSC17 score. b) Correlation with LSC 104 genes. Statistical differences were calculated between groups using a two-sided Student's t-Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant. c) Selected GSEA plots showing enrichment for HSC or LSC signatures.

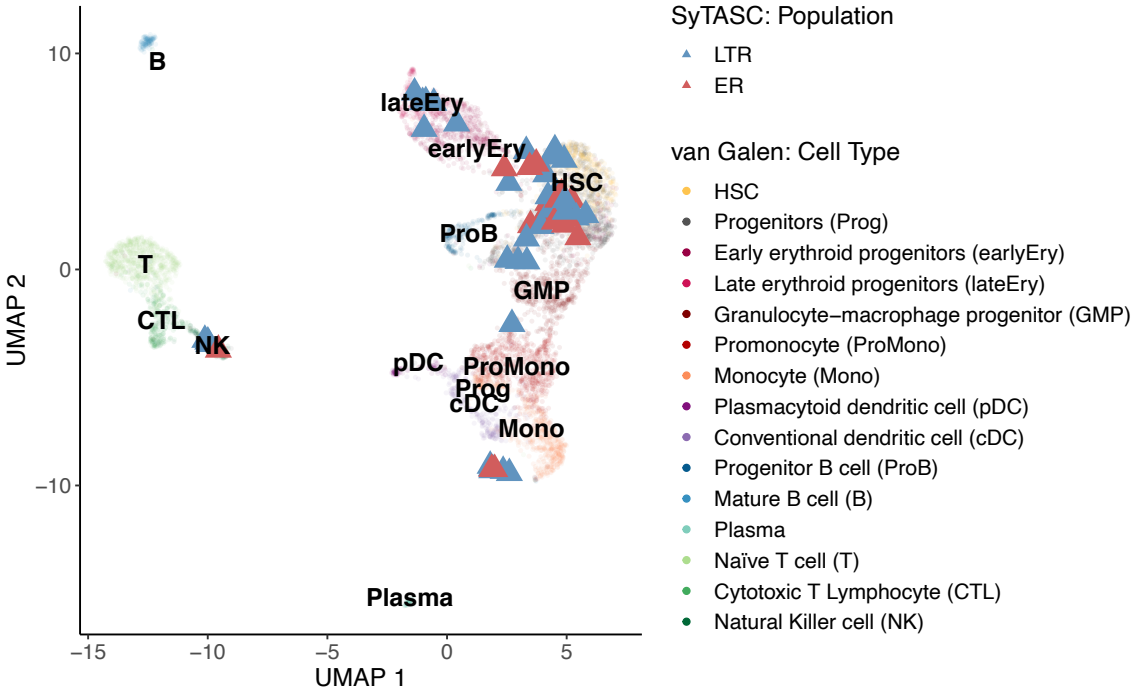

Figure 30: Embedding of LSC-enriched populations into a UMAP of single-cell data published by van Galen et al. as a reference.[40]

The more stem-like phenotype of ER samples was also reflected in unsupervised integration of the data sets by MOFA (cf. Figure 7). GSEA was performed for the loadings of LF3 and LF10 associated with the outcome groups. As shown in Figure 31a, ER samples were characterized by negative values of LF3 and gene sets related to HSC and LSC signatures were enriched in negative loadings of LF3. On the contrary, positive loadings of LF3 were associated to LTR. Positively enriched gene sets for LF3 were related to myeloid differentiation and development. (Figure 31a,b). Hence, the more stem-like phenotype in ER samples was reflected in multiple analyses.

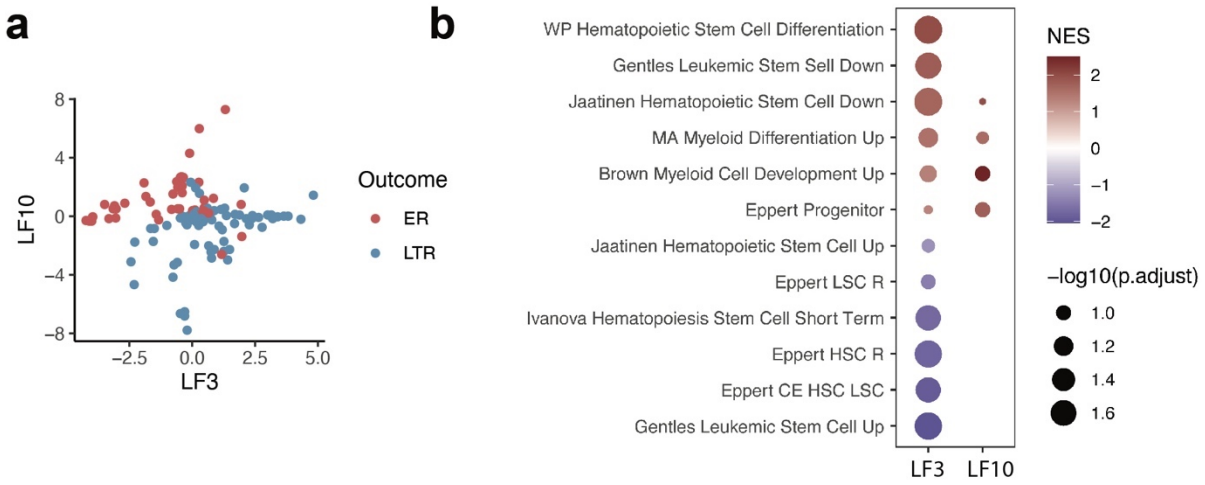

Figure 31: Association of MOFA LFs and outcome groups. a) Scatterplot showing loadings of LF3 and LF10. b) Bubble plot showing selected gene sets enriched when running GSEA on loadings of LF3 or LF10.

2.3.5 Potential effect of *DNMT3A* mutation on methylation and transcriptomic stability

As described above, the outcome groups displayed different global methylation patterns (cf. Figure S 10). Analysis of methylation patterns was particularly of interest since the mutation of *DNMT3A*, present in all patients, has been described to cause hypomethylation of DNA.[62] Comparison between the outcome groups identified 120 differential methylated regions in the LSC-enriched populations which were almost exclusively hypomethylated in ER samples. Accordingly, analysis of global methylation patterns between the outcome groups revealed lower beta values in ER samples; this effect was mainly driven by shore regions (Figure 32). Methylation patterns are highly correlated to the differentiation stages of cells in general (cf. Figure 14).[108] Therefore, differential methylation between the outcome groups might represent the differentiation stages of the LSC-enriched populations, or could reflect different cells of origin initiating the disease. However, the observation that ER samples had a more

stem-like phenotype is not in line with the global hypomethylation in this outcome group. Additionally, methylation was also lower in the more differentiated populations and clearly lower in AML compared to heathy samples (Figure S 11). This was particularly observed in the most differentiated populations ($CD34^{+}GPR56^{+}NKG2DL^{+}$ and $CD34^{-}GPR56^{-}NKG2DL^{+}$) that seemed to be phenotypically similar between healthy and AML based on RNA-seq (cf. Figure 8a, Figure 12, and Figure S 3). Hence, the *DNMT3A* mutation induced a global hypomethylation in this cohort. These observations led to a second hypothesis, that the *DNMT3A* mutation might have occurred earlier in the ER samples than in the LTR, leading to a more disturbed methylome, transcriptome and thus a more aggressive type of AML.

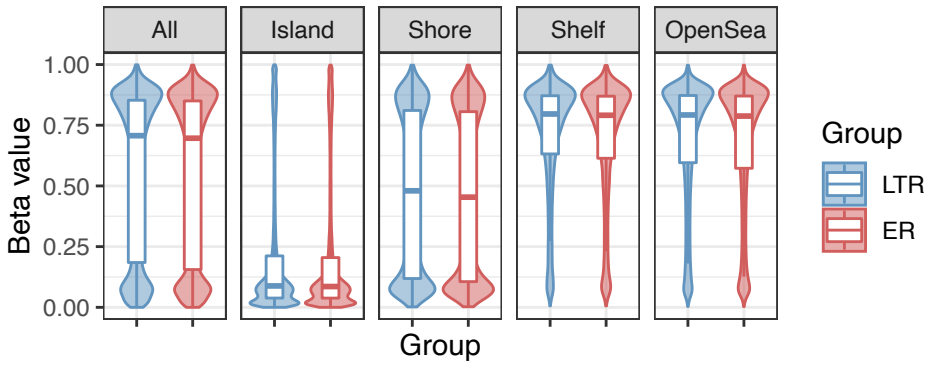

Figure 32: Distribution of beta values across all regions, island, open sea, shelf, and shore regions for ER and LTR LSC-enriched populations.

Alterations congruent with increased transcriptomic instability were also observed in an analysis of alternative splicing. Alternative splicing events were very frequent in the LSC-enriched populations, although sufficiently recurrent coordinated differences of single events between the two outcome groups with inclusion level differences > 50% were not present (Figure 33). However, in total, a lower number of splicing events was observed in the ER samples. Particularly comparing outcome groups across all populations, the number of intron retention events was higher, whereas the number of exon skipping was lower in the ER group (Figure 33 and Table S 2). Comparing differentially methylated and alternatively spliced genes also indicated co-occurence of these processes (Table S 3). This further supported the hypothesis that the mutation of *DNMT3A* caused progressive hypomethylation that might lead to increased transcriptomic instability in ER samples, possibly by an earlier first mutational event in these samples and a longer time lapse before the second genetic hit, i.e., the *NPM1* mutation.

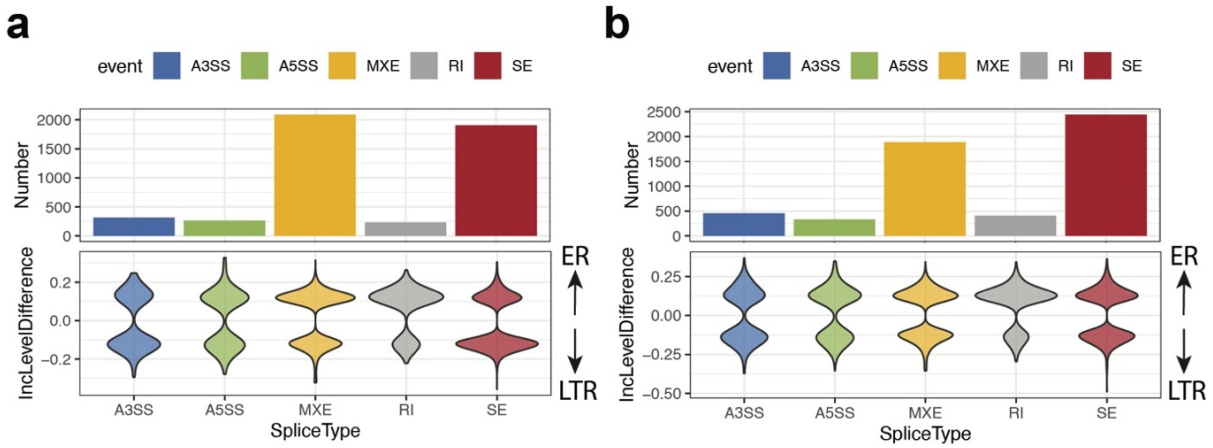

Figure 33: Quantification of alternative splicing events between ER and LTR. a) All sorted populations. b) Engrafting LSC-enriched populations. RI: intron retention, MXE: mutually exclusive exons, SE: exon skipping, A5SS: alternative 5' donor site, A3SS: alternative 3' donor site. Statistics was filtered for events with an absolute inclusion level difference > 0.1 and an FDR < 0.05.

2.3.6 Higher *NPM1* mutant allele frequencies in ER samples

Besides mutated *DNMT3A*-p.R882, the cohort harbors an *NPM1*-p.W288fs*12 mutation. While *DNMT3A* displayed no difference of mutant allele frequencies, the frequency of *NPM1*-mutant alleles was significantly higher in ER compared to LTR samples (Figure 34). This trend could not fully be explained by the abundance of healthy retained HSCs in the $CD34^{+}GPR56^{+}NKG2DL^{-}$ population as described above (cf. subsection 2.2.5). The frequency for those samples, for which the type was determined, was particularly low in samples that showed multi-lineage. Interestingly, multi-lineage engraftment occurred more often in LTR compared to ER samples; particularly, in the unsorted bulk samples (Table S 1). Hence, the higher allele frequency might indicate an advanced AML type characterized by a more dominant *NPM1*-mutated clone.

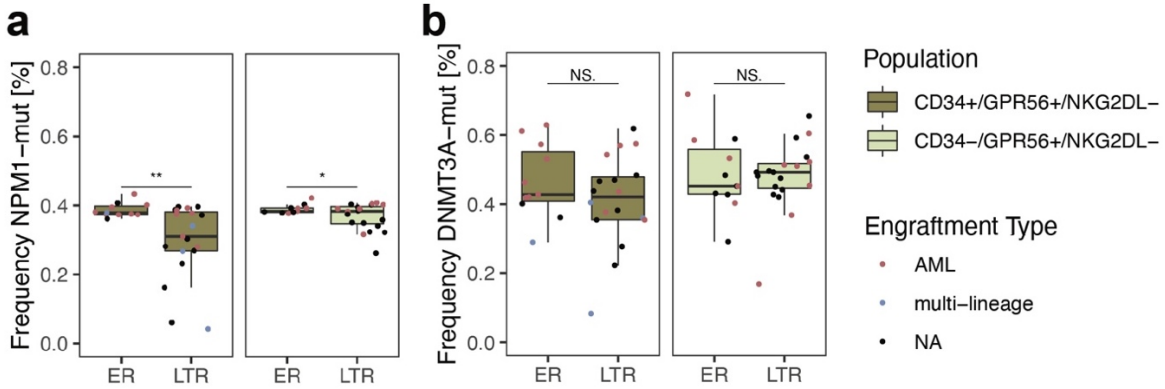

Figure 34: Mutant allele frequency based on RNA-seq data between ER and LTR for LSC-enriched populations. a) Percentage of *NPM1*-mutated counts. b) Percentage of *DNMT3A*-mutated counts. Statistical differences were calculated between groups using a two-sided Student's t-Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

2.3.7 Alteration of energy metabolism in engrafting LSC populations

To understand underlying biological processes driving the differential outcome between ER and LTR samples in the engrafting LSC populations, differential expression analysis followed by GSEA was performed. Several dominant alterations were related to energy metabolism. As shown in Figure 35a, gene sets related to glycolytic processes were significantly enriched in LTR samples. This effect was more pronounced in the CD34^{+} LSC population. The enrichment was mainly driven by differentially expressed genes highlighted in Figure 35b; for example, Pyruvate Carboxylase (*PC*), Pyruvate Kinase L/R (*PKLR*), or Hexokinase 3 (*HK3*).

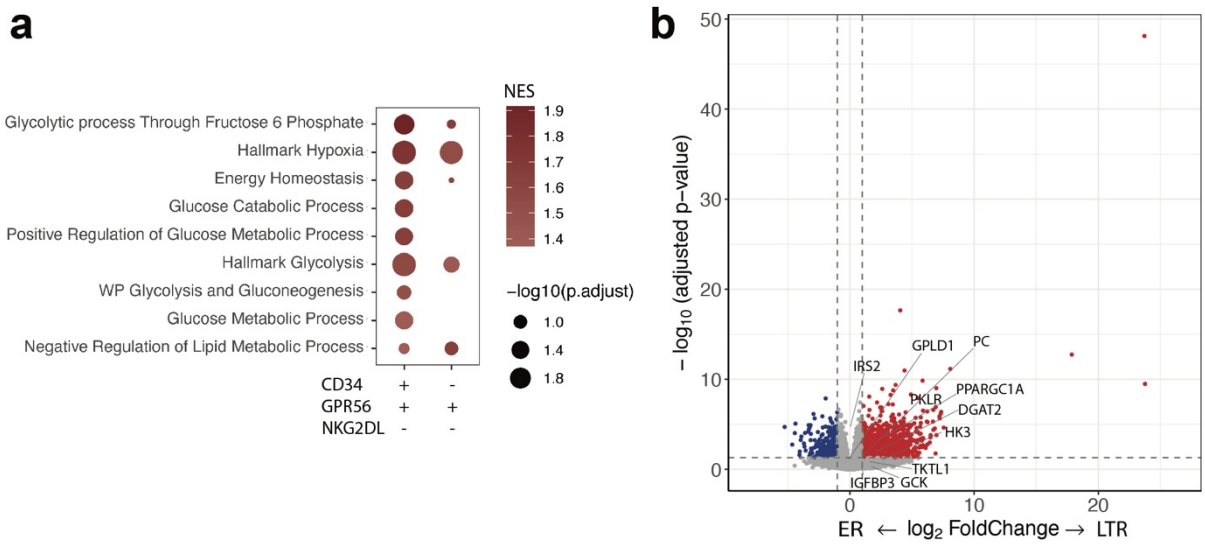

Figure 35: Enrichment of glycolysis-related gene sets in LTR samples. a) GSEA for the two LSC populations separately. A positive NES indicates enrichment in LTR samples. b) Volcano plot highlighting genes which are driving the enrichment of glycolysis-related processes based on leading-edge analysis. Differential expression statistics indicate the comparison between ER and LTR samples for both LSC-enriched populations combined.

In contrast to LTR samples, mitochondrial genes involved in the respiratory chain complexes 1,4, and 5 were overexpressed in ER samples indicating a higher activity of oxidative phosphorylation. Interestingly, gene sets related to oxidative phosphorylation, oxidative stress, and the TCA cycle also known as the citric acid cycle were enriched in the bulk data set even though statistical significance was lacking (Figure S 12).

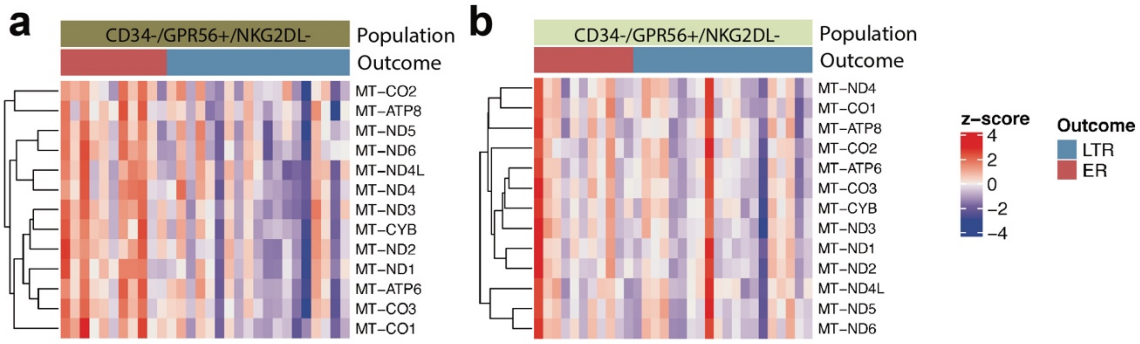

Figure 36: Expression of mitochondrial genes involved in the respiratory chain complexes 1,4,5. a) Heatmap for CD34^{+}GPR56^{+}NKG2DL^{-} LSC population. b) Heatmap for CD34^{-}GPR56^{+}NKG2DL^{-} LSC population.

To investigate this observation in more detail, metabolomics analysis of CD34^{+}GPR56^{+} and CD34^{-}GPR56^{+} populations from 3 ER and 3 LTR samples, derived from patient-derived xenograft (PDX) experiments, was performed. NKG2DL was not included in the sorting strategy. However, the samples processed by Dr. Elisa Donato were derived from engrafting PDX samples and thus mature human cells were not expected. Hence comparable populations (CD34^{+}GPR56^{+}NKG2DL^{-} and CD34^{-}GPR56^{+}NKG2DL^{-}) as used for RNA-seq and methylation could be assumed. The untargeted metabolomics measurements were performed by Prof. Nicola Zamboni and allowed quantification of 344 metabolites covering different metabolic pathways as presented in Figure S 13.

Between ER and LTR, 4 and 15 metabolites were observed to be significantly differentially abundant, in the CD34^{+}GPR56^{+} and CD34^{-}GPR56^{+} populations respectively. Hexose bisphosphate levels were higher in ER in both sorted populations (Figure 37a,b). While oxoglutarate (α-ketoglutarate) was significantly more abundant in LTR in the CD34^{+}GPR56^{+} population, intermediates of anaplerotic reactions replenishing the TCA cycle (e.g., urocanate) showed significantly higher concentrations in LTR in the CD34^{-}GPR56^{+} population. Of note, the enrichment of hydroxyglutarate was mainly driven by SyT50, an IDH1-mutant patient sample. The abundance of other metabolites such as succinate was similar to the other LTR samples (Figure S 14 and cf. Figure 5). Hence, hydroxyglutarate was removed from the pathway enrichment analysis. Pathway enrichment partially reflected the differentially abundant metabolites. For example, amino acid metabolism (e.g. alanine, glutamate, histidine, tyrosine) was enriched in LTR (Figure 37c and cf. Figure 5). This observation could be reconciled with the hypothesis of altered energy metabolism. However, due to the low number of differentially abundant metabolites, clear conclusions could not be drawn. In

particular, the enrichment of pathways in LTR samples in the CD34^{+}GPR56^{+} population was only based on a few metabolites.

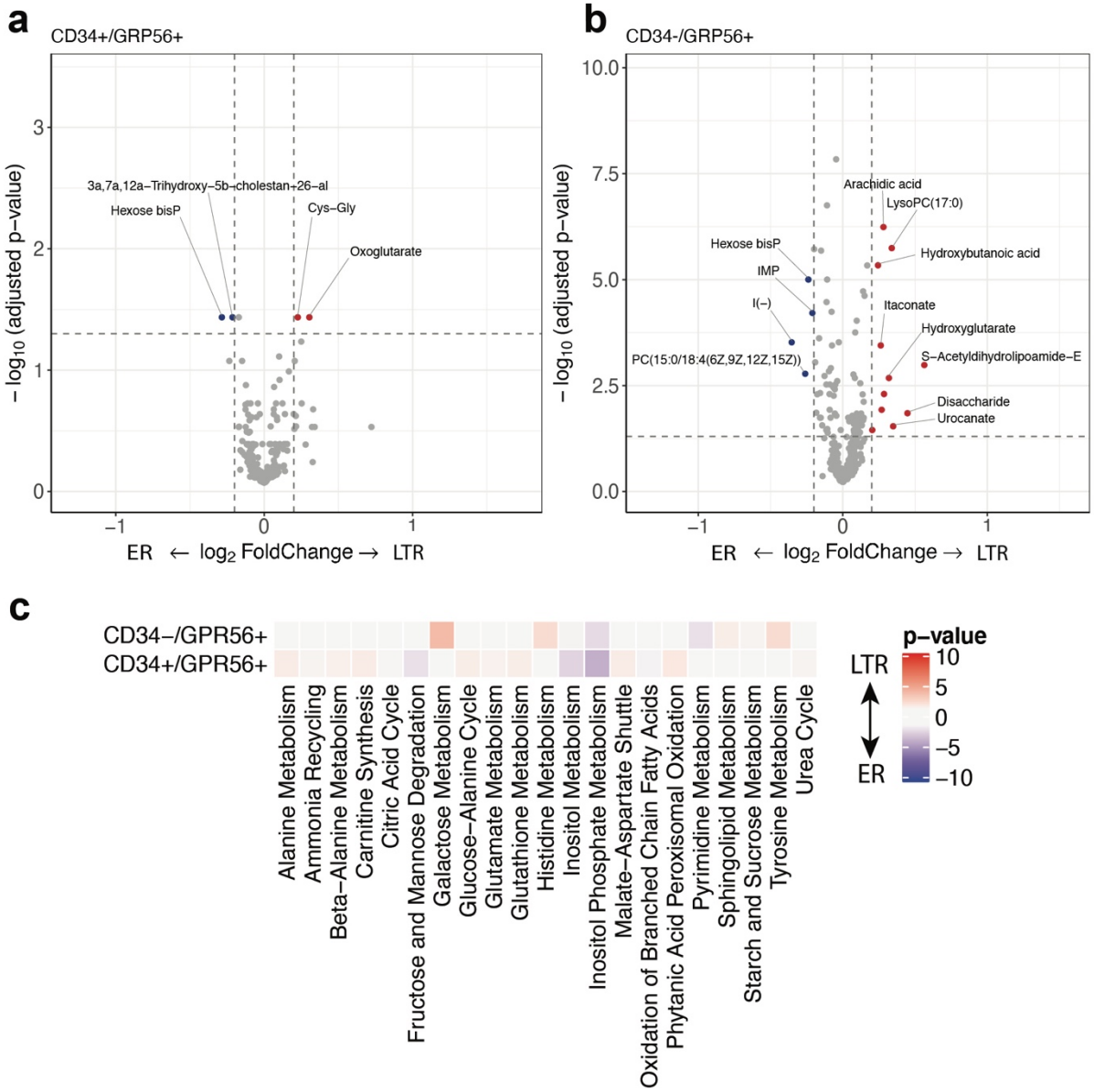

Figure 37: Metabolic differences between outcome groups in CD34^{+}GPR56^{+} and CD34^{-}GPR56^{+} populations. a) Volcano plot for CD34^{+}GPR56^{+} populations. b) Volcano plot for CD34^{-}GPR56^{+} populations. An absolute $\log_2$ fold change > 0.2 was considered differential. c) Heatmap showing the significance of enriched metabolic pathways between ER and LTR.

While most pathways showed enrichment for either of the populations, inositol phosphate metabolism was significantly enriched in ER samples for both populations (Figure 37c). This was in line with the analysis of key pathways based on RNA-seq data as shown in Figure 38. PI3K signaling was clearly more active in ER samples. Thus, metabolomics and RNA-seq data showed activity of similar biological processes.

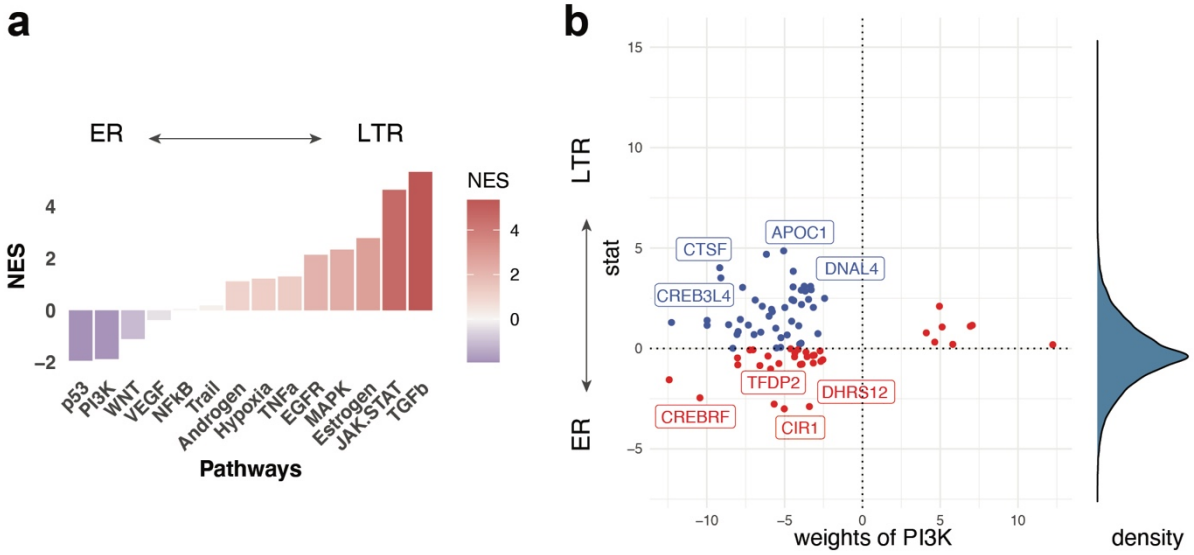

Figure 38: Activity of key pathways inferred by PROGENy between outcome groups in engrafting LSC-enriched populations. a) NES for key pathways. b) Exemplary expression and weights of genes driving the activity of the PI3K pathway.

2.3.8 Increased TGFβ signaling in LTR samples

The analysis of key pathway activity revealed overall higher enrichment levels in LTR samples. TGFβ signaling was enriched strongest in this outcome group (Figure 38a). The activity of TGFβ signaling was mainly driven by ID4 (Inhibitor of Differentiation 4) which was one of the most significantly differentially expressed genes (Figure 39a,c). Within the cellular signaling network, this pathway shares signaling cascades with the bone morphogenetic protein (BMP) pathway as BMPs are part of the TGFβ family.[190] GSEA showed strong enrichment for of TGFβ- and BMP-related gene sets for LTR samples in both LSC-enriched populations (Figure 39b). Interestingly, Noggin (*NOG*), which is an inhibitor of BMP signaling, was one of the most differentially overexpressed genes in LTR compared to ER samples (Figure 39c).

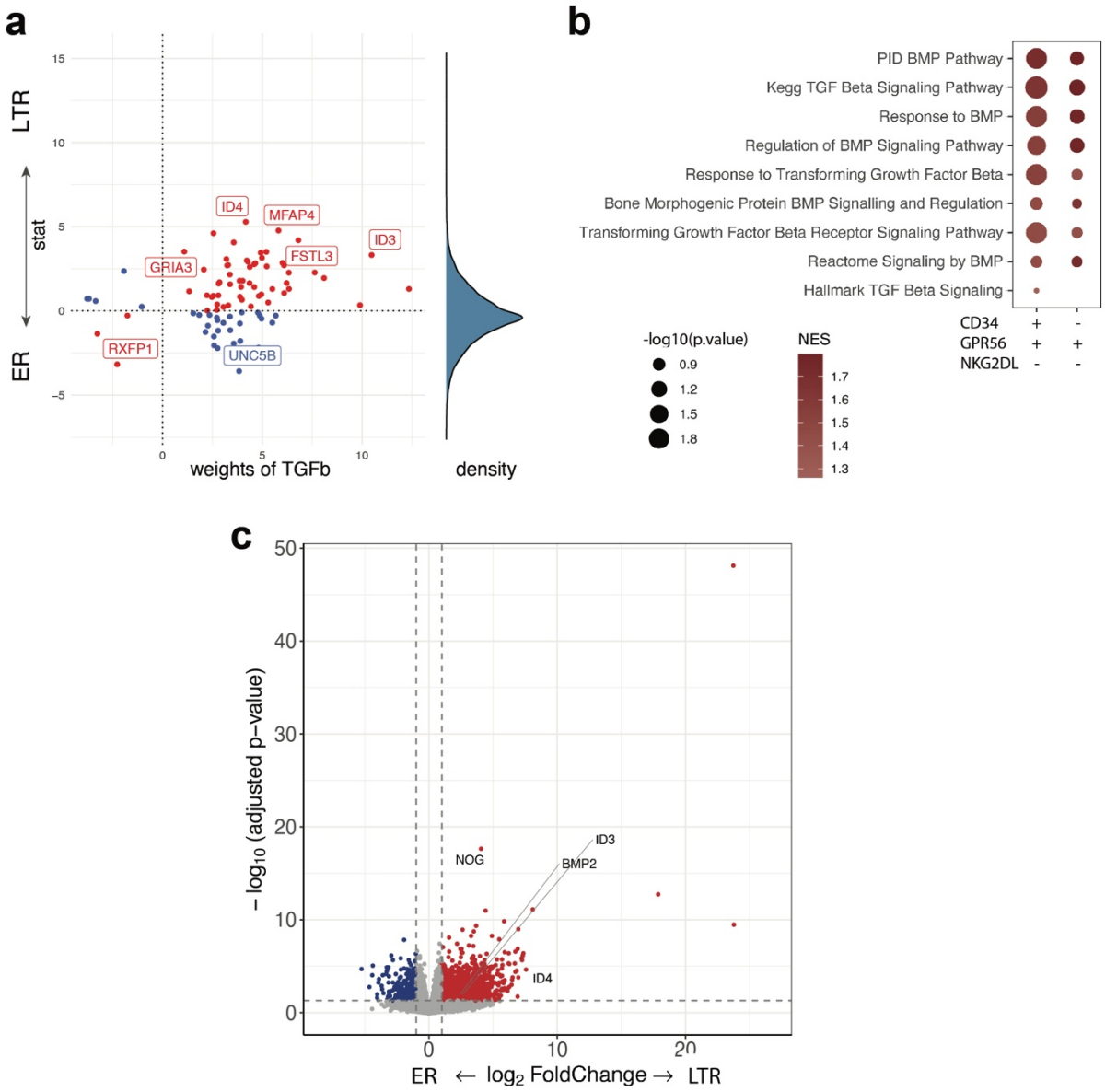

Figure 39: Enrichment of BMP and TGFβ signaling in LTR LSC-enriched populations compared to ER samples. a) Exemplary expression and weights of genes driving the activity of the TGFβ pathway inferred by PROGENy. b) GSEA for the two LSC populations separately. A positive NES indicates enrichment in LTR samples. c) Volcano plot highlighting genes driving the enrichment of BMP and TGFβ-signaling based on leading-edge analysis. Differential expression statistics indicate the comparison between ER and LTR for both LSC populations combined.

2.3.9 Trained outcome prediction signature is highly predictive in external AML cohorts

The RNA-seq data was used to train an outcome prediction signature via LASSO regression with ER and LTR samples as two groups. Technical details are documented in the materials and methods chapter.

As expected, the trained 30-gene-signature showed high predictive power in its own training cohort when the median signature score was used for stratification into “low” and “high” groups associated with “good” and “poor” prognosis, respectively (Figure 40). Accordingly, the signature genes were clearly differentially expressed between the outcome groups (Figure 41a and Table S 4). Over-representation analysis revealed that only few of the signature genes shared common biological functions or were specific for certain biological processes. For example, four genes shared enrichment for regionalization or pattern specification (Figure 41b). Interestingly, some of the signature genes also showed significantly differential methylation in their promoter region, e.g. Sortilin 1 (*SORT1*) and Ectonucleotide Pyrophosphatase/Phosphodiesterase 2 (*ENPP2*) between the outcome groups in the LSC-enriched populations (Figure 42).

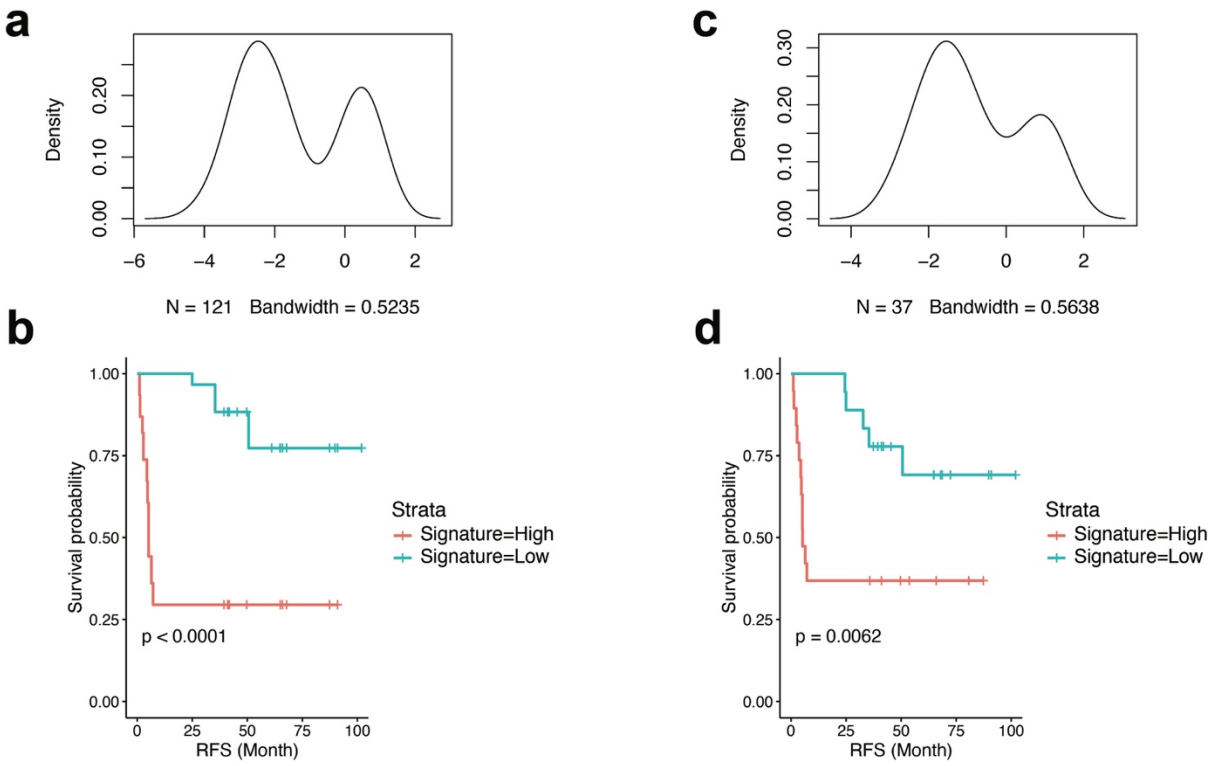

Figure 40: Distribution of signature scores and Kaplan-Meier curves for training cohort. a,b) Results for population-sorted data set. c,d) Results for bulk data set. P-values were calculated using a log-rank test on the groups stratified by the median of the respective signature scores in cohorts.

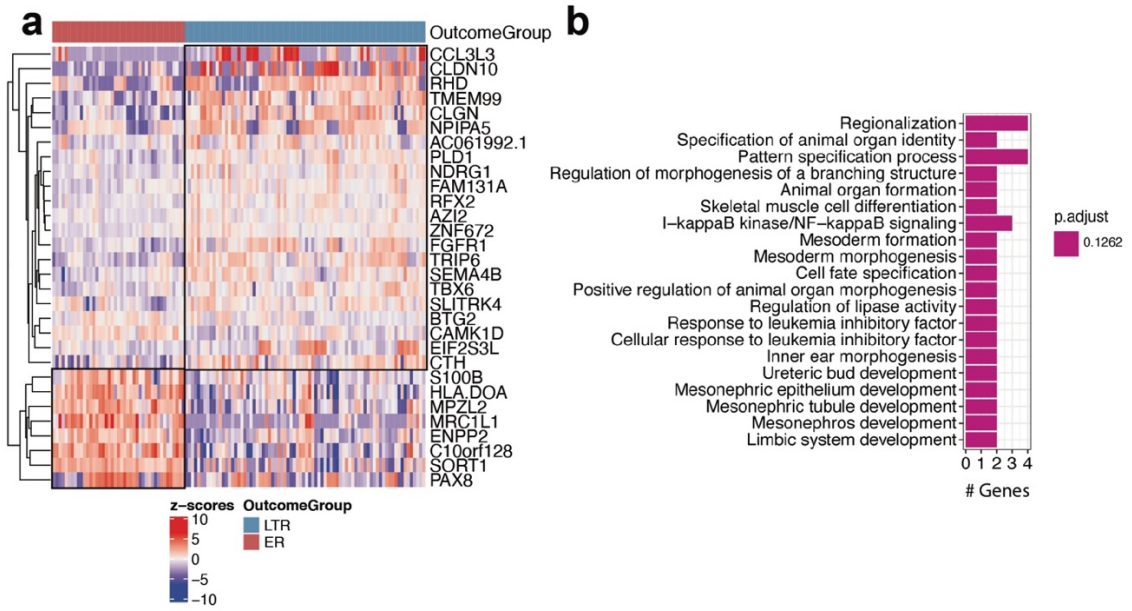

Figure 41: Expression and biological function of signature genes. a) Heatmap showing the expression of signature genes stratified by the outcome groups in the population-sorted data set. Black frames indicate upregulation in either ER or LTR samples. b) Bar plot showing over-representation analysis of signatures genes.

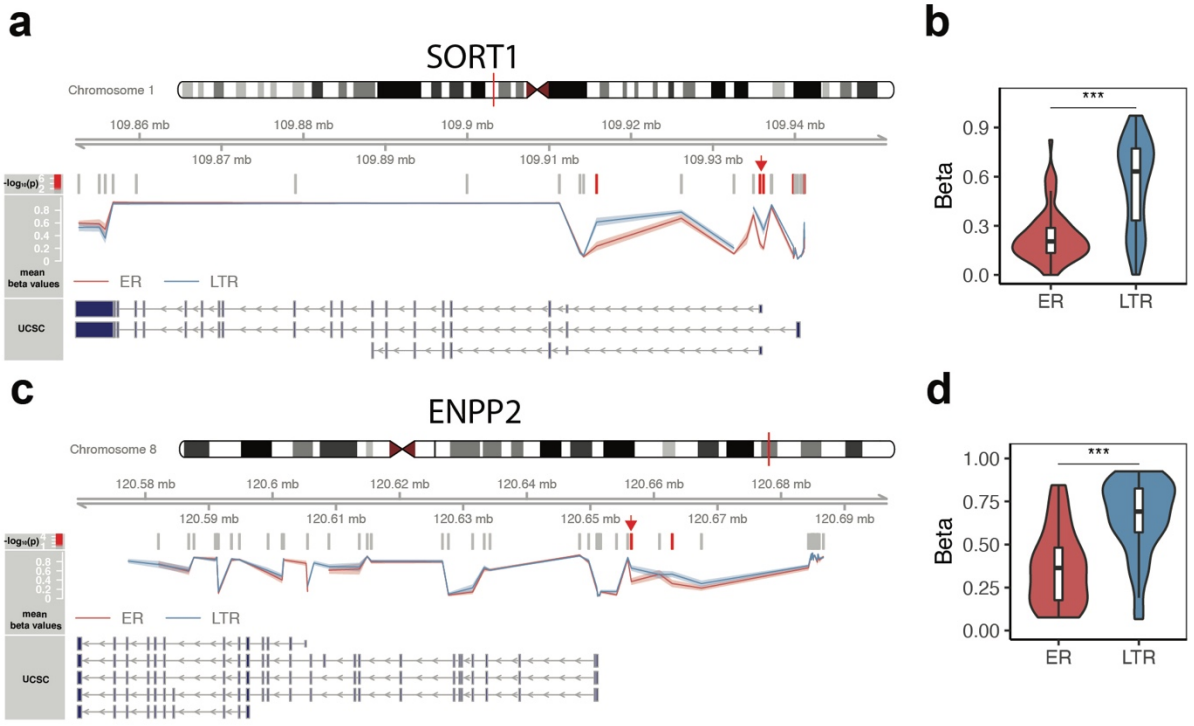

Figure 42: Methylation of signature genes. Positions differentially methylated between engrafting populations of ER and LTR cohorts. q-values were calculated via differentially methylated position analysis. UCSC exonic regions in the bottom track correspond to genome version hg19. a,b) Methylation of *SORT1*. c,d) Methylation of *ENPP2*. Positions used for quantification of beta values as shown in violin-box plots in (b) and (d) are marked by red arrows in (a) and (c). Statistical differences were calculated between groups using a two-sided Wilcoxon Rank Sum Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

To test whether the signature was also predictive in other AML cohorts, scores were calculated for six data sets and each cohort was stratified based on the respective median signature score. Since the RFS information was sparse for most cohorts, the overall survival (OS) was used for statistical testing. For the TCGA AML data sets, the signature showed high predictive power. Of note, the TCGA AML data set was downloaded from two different repositories referred to as “TCGA” and “GDC TCGA” since the documented survival data and preprocessed data were not identical in these repositories.[59] Filtering for normal karyotype AMLs generally increased statistical significance. For the “TARGET “ data set only a trend toward correlation of high signature score and poor prognosis was observed; while the score showed significant predictive power after filtering of the “Beat (OSHU)” cohort (Figure 43a).[90,191] The signature prediction was also highly significant for the “Oellerich” cohort, while only a tendency was observed for the “Metzeler” microarray data set (Figure 43b).[192,193] In summary, the trained signature showed high predictive power also in external AML cohorts.

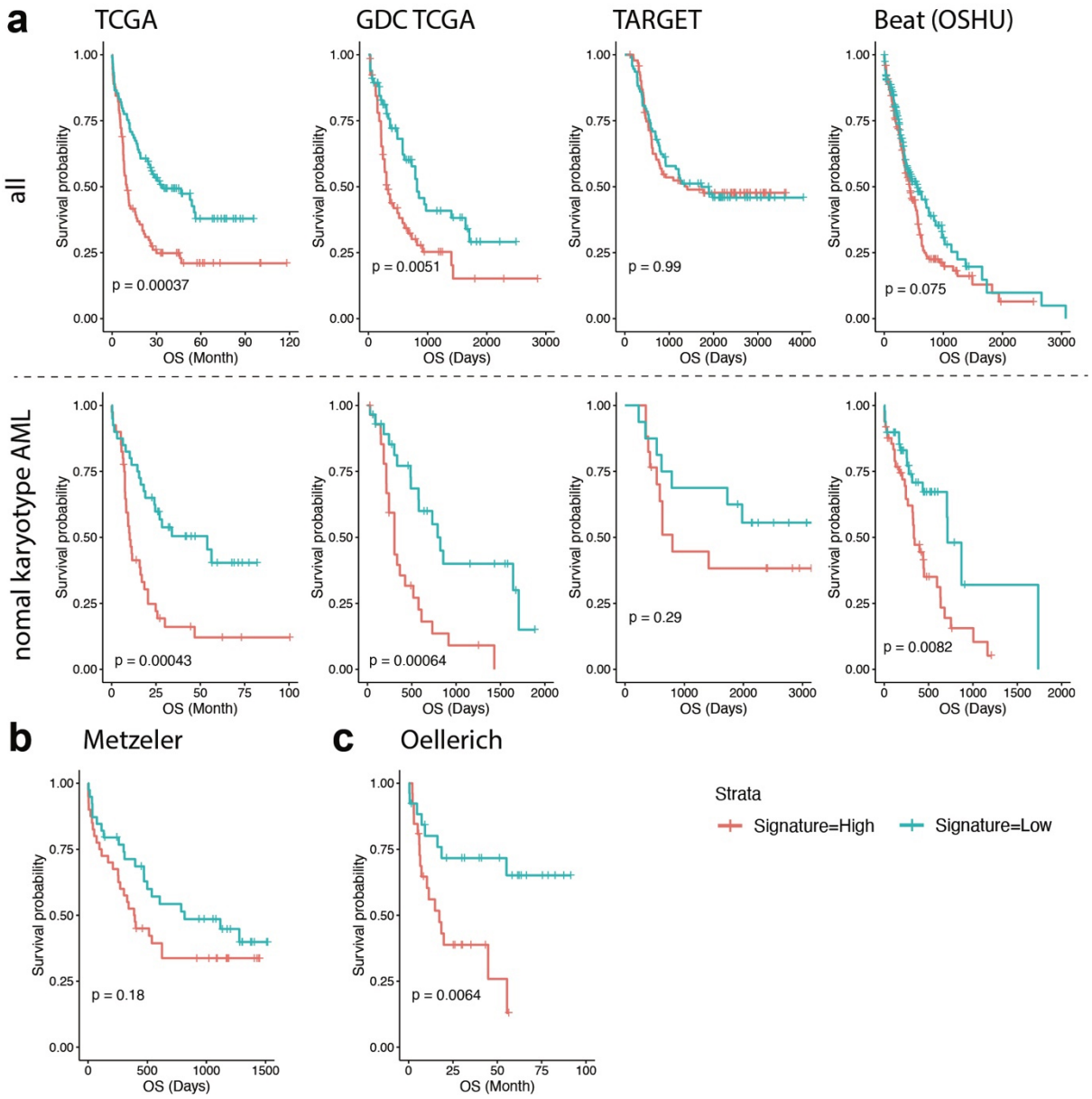

Figure 43: Kaplan-Meier curves for external cohorts. a) Curves for TCGA AML, GDC TCGA AML, TARGET AML, and the Beat (OSHU) cohort. The top row includes all AML samples. The bottom row was filtered for AML samples presenting a normal karyotype. b) Curve for the Metzeler cohort. c) Curve for the Oellerich cohort. P-values were calculated via a log-rank test on the groups stratified by the median of the respective signature scores in cohorts. Color code represents these strata as indicated in the bottom right corner applying to all plots.

To estimate the performance of the signature, the LSC17 score was analogously calculated for the external cohorts and the predictive power was assessed. The LSC17 score is a signature used to estimate the risk potential of AML.[51] Comparison of the trained signature with the LSC17 results showed mixed results between the cohorts. While the TCGA cohorts and the "Oellerich" cohort were slightly more significant for the trained outcome prediction signature, the LSC17 performed strikingly better for the "TARGET" cohort and slightly better for the "Beat (OSHU)" cohort. The prediction for the "Metzeler" cohort was not significant for the outcome prediction signature but showed reversed strata for the LSC17 score (Figure 44 and Figure 45). Accordingly, the predictive power of the outcome signature was comparable, if not better compared to the LSC17 score.

The LSC17 score was trained to predict malignant stemness in AML cohorts which is an indicator of prognosis.[51] The outcome prediction signature, however, was trained using the time to relapse for samples taken at diagnosis. Consequently, I combined both scores to investigate whether the predictive power was additive. The median score was used for each cohort and signature separately to stratify the data sets into four groups; either "high" or "low" for both LSC17 and outcome prediction signature. For all cohorts, predictive power indicated by median survival time was increased when both signatures were combined, as was statistical significance; except for the "Beat (OSHU)" cohort (Figure 44 and Figure 45). The same applied to the combination of the outcome prediction signature with the ELN classification approach based on molecular risk ("good", "intermediate" or "poor"). This classification was only available for the "TCGA" cohort and itself displayed high predictive power (Figure 46a). Combination with the outcome prediction signature was implemented analogously to the described procedure yielding six different strata. Again the median survival time showed increased discriminatory power for all risk groups and particularly separated the "intermediate" risk group with high statistical significance (Figure 46b). Taken together, I could show an additive effect of the trained outcome prediction signature with the established LSC17 score and the ELN classification.

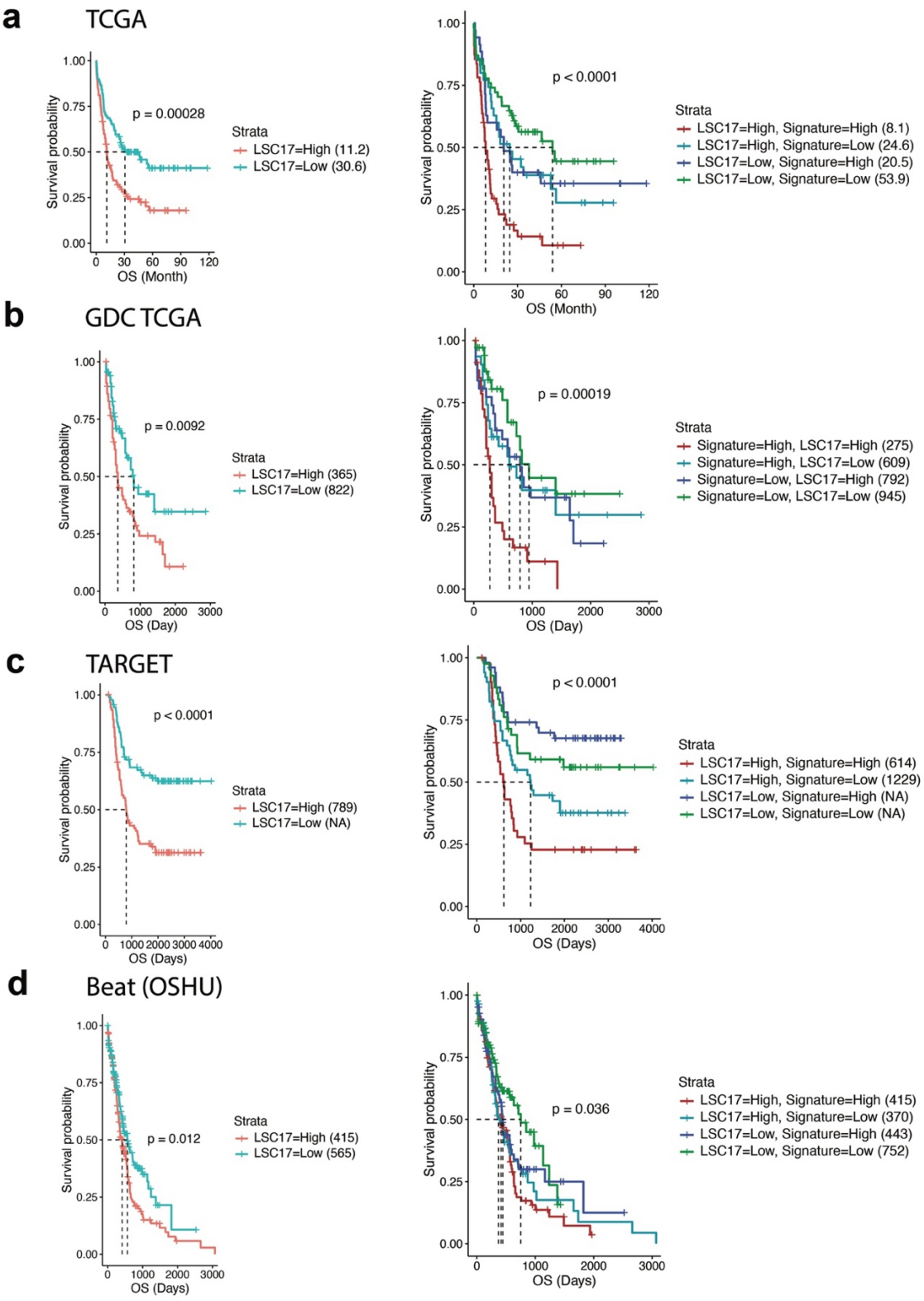

Figure 44: Kaplan-Meier curves for external cohorts combining the trained signature and the LSC17 score. Left plots show Kaplan-Meier curves for LSC17 scores. Right plots show combinations of both signature scores. a) TCGA AML. b) GDC TCGA AML. c) TARGET AML. d) Beat (OSHU) cohort. P-values were calculated via a log-rank test on the groups stratified by the median of the respective signature scores in each cohort. The number in parentheses for each stratum indicates the median survival time corresponding to the dashed line in each plot.

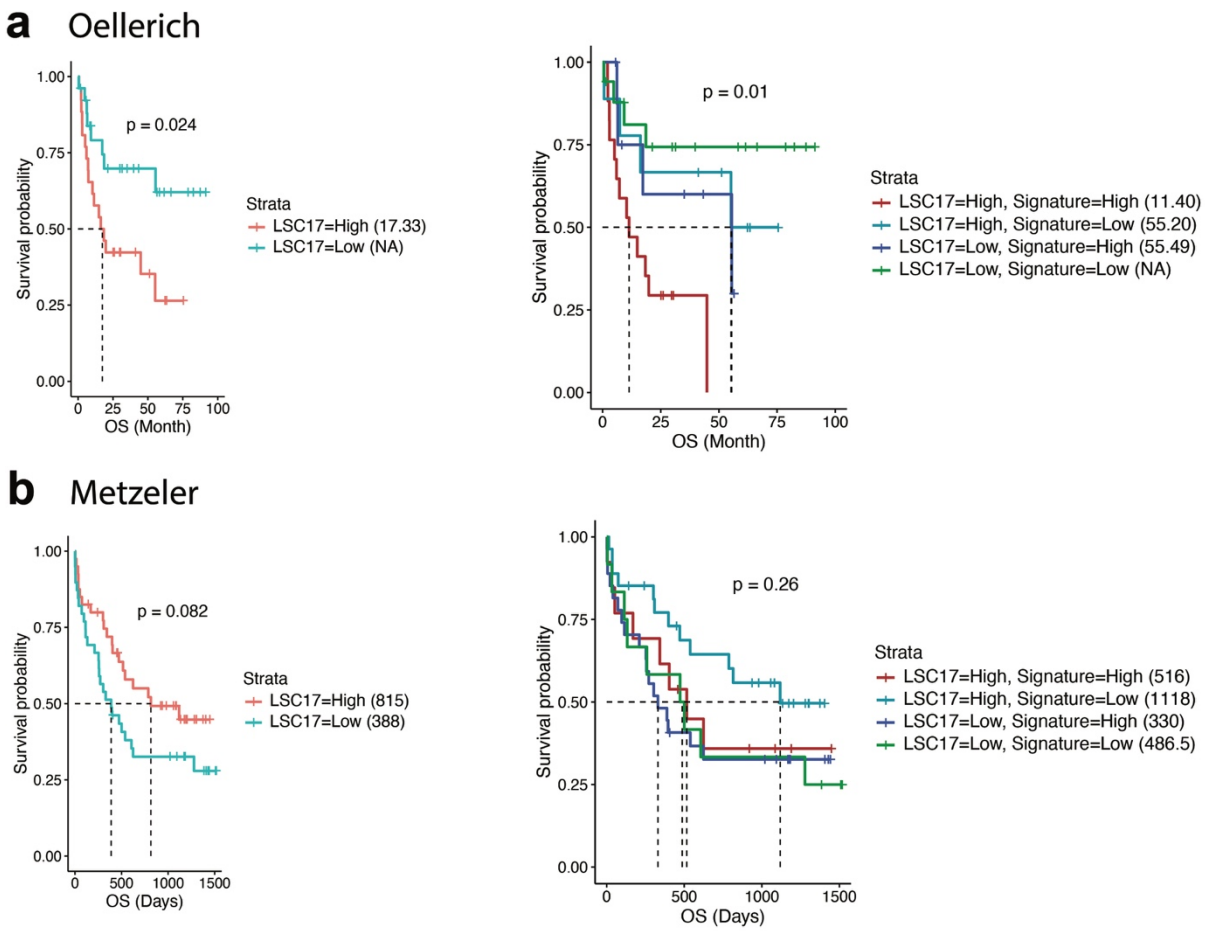

Figure 45: Kaplan-Meier curves for external cohorts combining the trained signature and the LSC17 score. Left plots show Kaplan-Meier curves for LSC17 scores. Right plots show combinations of both signature scores. a) Oellerich. b) Metzeler. P-values were calculated via a log-rank test on the groups stratified by the median of the respective signature scores in each cohort. The number in parentheses for each stratum indicates the median survival time corresponding to the dashed line in each plot.

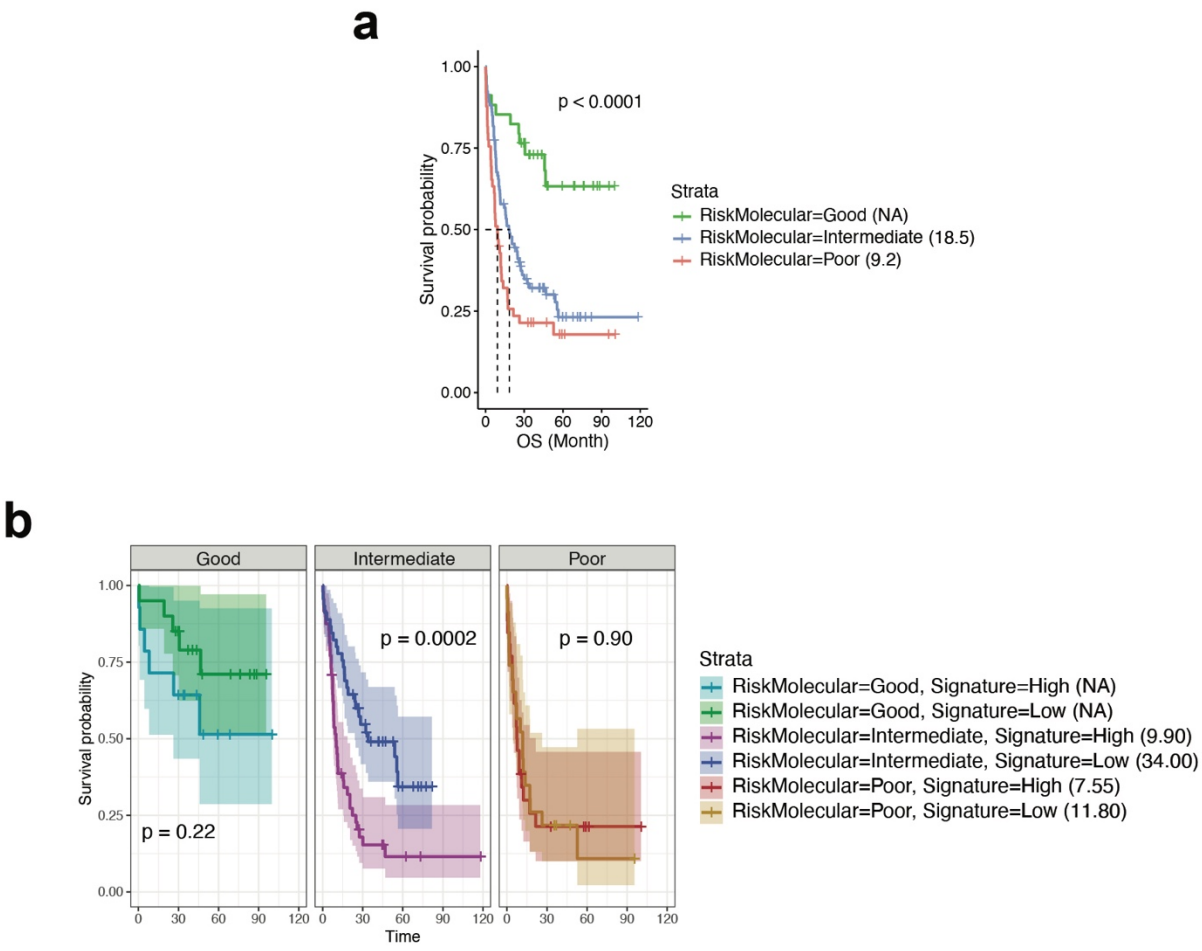

Figure 46: Kaplan-Meier curves for the TCGA cohort combining the trained signature and the molecular risk. a) Kaplan-Meier curve for molecular risk determined by the ELN classification scheme as documented by TCGA. b) Kaplan-Meier curves combining molecular risk and trained signature score. P-values were calculated via a log-rank test on the groups stratified by the median of the respective signature scores in each cohort. The number in parentheses for each stratum indicates the median survival time corresponding to the dashed line (only plot a)).

2.3.10 High expression of MHC-II in ER samples

One of the signature genes with highest coefficients (i.e., predictive power for an early relapse) was *HLA-DOA*, an MHC-II gene (Figure 41a). MHC-II gene sets were also identified as highly enriched in ER compared to LTR samples across all populations (Figure 27c). Direct comparison between the outcome groups for engrafting LSC-enriched populations also revealed striking enrichment for MHC-II-related gene sets in ER samples. Again, this effect was more pronounced in the $CD34^{+}GPR56^{+}NKG2DL^{-}$ population (Figure 47a,b). Analysis of transcription factor activity showed that the differential expression was likely regulated it by RFX5, RFXANK, and RFXAP which are known to activate transcription in MHC-II promoters (Figure 48).[194]

Notably, MHC-II genes were also differentially expressed in the bulk data set (Figure S 17). Taken together, MHC-II expression and antigen presentation were identified as one of the most strongly altered biological processed across all cell populations.

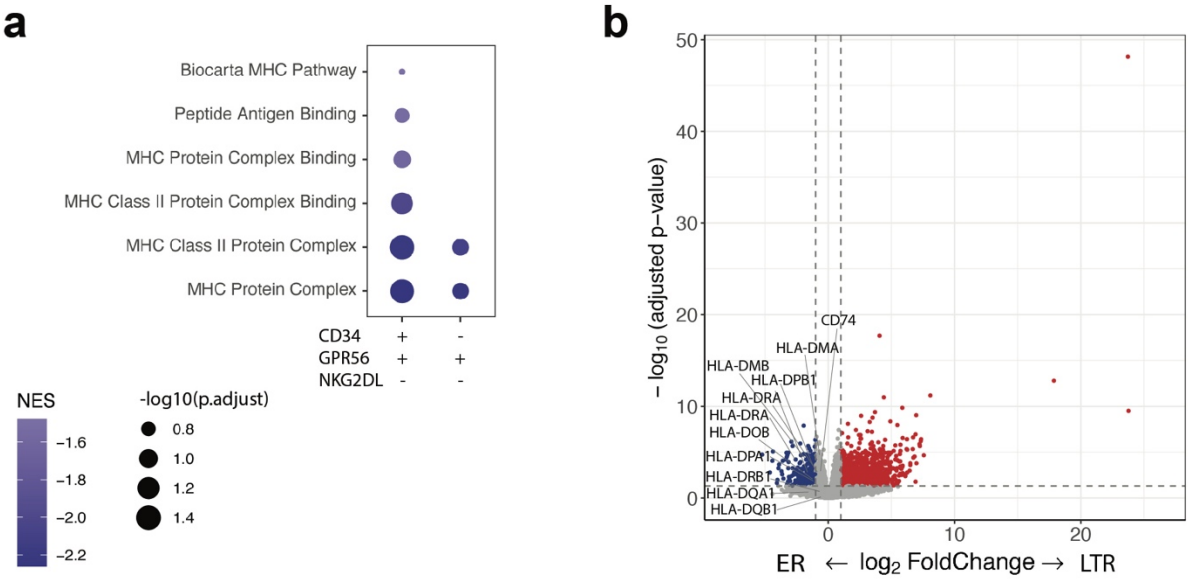

Figure 47: Enrichment of MHC-II genes in differentially expressed genes between ER and LTR LSC-enriched populations. a) Bubble plot of enriched gene sets related to MHC-II expression by GSEA. Statistics are shown separately for $CD34^{+}$ and $CD34^{-}$ LSC-enriched populations. b) Volcano plot of differentially expressed genes highlighting MHC-II genes.

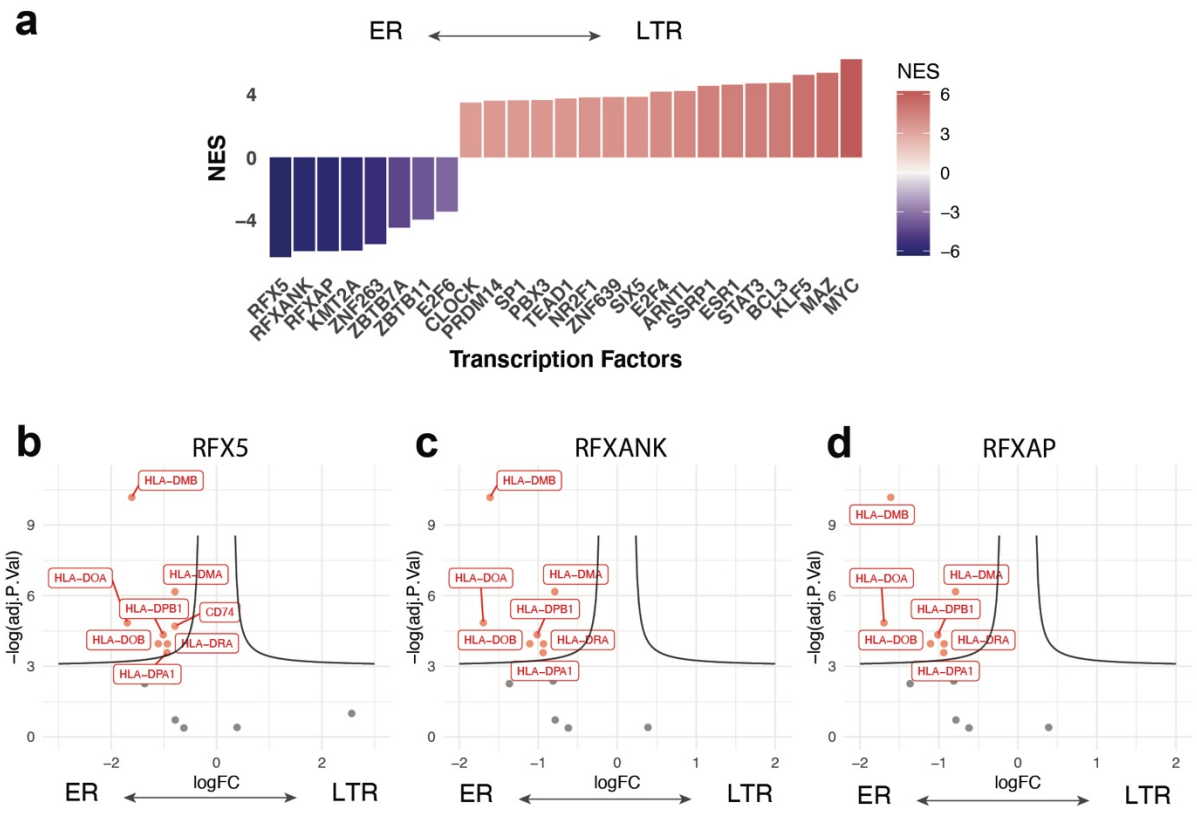

Figure 48: Activity of transcription factors inferred by VIPER between ER and LTR LSC populations. a) NES for 25 most differentially active transcription factors. Exemplary volcano plots for genes driving the enrichment for transcription factors b) RFX5, c) RFXANK, and d) RFXAP.

In summary, I could show that the SyTASC data set provides a homogenous genetic background with no major confounding factors. Differences based on RNA-seq and methylation data were more pronounced in the engrafting LSC-enriched populations. These biological differences included changes in key pathways such as TGFβ/BMP and PI3K signaling. Increased hypomethylation and analysis of the mutant allele frequencies indicated a higher transcriptomic instability which may be caused by an earlier occurrence of the *DNMT3A* mutation. In addition, differences in the energy metabolism were observed which could partially be confirmed in metabolomics analyses. In a second part I trained an outcome prediction signature that exhibited high predictive power in multiple AML cohorts and additive discriminatory power in combination with established classifiers.

2.4 Optimization of intracellular metabolomics measurements

Intracellular metabolite measurements require suitable protocols for their extraction. An optimal protocol allows efficient extraction of a wide range of metabolites as well as repeatability of results to ensure a low technical variation between samples and replicates. Independent from the SyTASC project, I worked on a comparative study to identify optimal extraction protocols for different human sample types. As outlined in Figure 49, this project included four sample types: liver, bone marrow, and two cell lines (adherent human embryonic kidney (HEK) and non-adherent human leukemia 60 (HL60)). These samples were used for intracellular metabolite extraction with ten different established protocols. Metabolite quantification was performed by a commercial kit (Biocrates MxP® Quant 500) covering a large range of up to 630 metabolites.

During this study as well as in exchange with collaboration partners, a need for an easy-to-use platform for customized analysis became apparent. Therefore, together with Nils Mechtel, I set up the R package "MetAlyzer" available on CRAN (https://CRAN.R-project.org/package=MetAlyzer). This tool facilitates the reading and processing of the standardized output data from the MetIDQ™ platform provided by Biocrates. Figure 50 illustrates further functionalities including data handling, statistics, and downstream analysis. The optimal extraction protocol for a study depends on multiple factors such as sample type and metabolites of interest. Therefore, Nils Mechtel and I made the data set available as an interactive R Shiny app "MetaboExtract" (http://www.metaboextract.shiny.dkfz.de). This app allows users to explore and compare extractions protocols. Tissues, extraction protocols and metabolite classes can be (de)selected to focus on the data of interest. Additionally, the limit of detection (LOD) can be used to filter the data and the maximal coefficient of variation (CV) between replicates can be selected by users to identify the most suitable method.
The following subsection describes the study results with default filtering options. As an application example, I outline the rationale for selecting the extraction method used for pre-processing the untargeted metabolomics study for the SyTASC samples as presented in subsection 2.3.7.

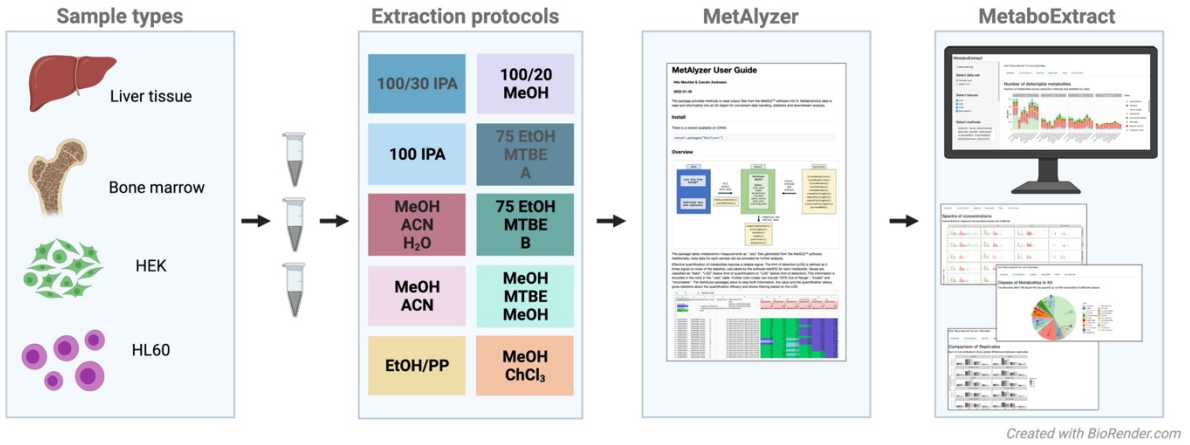

Figure 49: Workflow of the comparative metabolomic study and associated analytic software. Four human sample types were extracted in triplicates using ten different extraction protocols. Metabolomics output was read and processed using the package "MetAlyzer", programmed by Nils Mechtel and me. Statistics and data were made available online as the R Shiny app "MetaboExtract", also programmed by Nils Mechtel and me, for the external user.

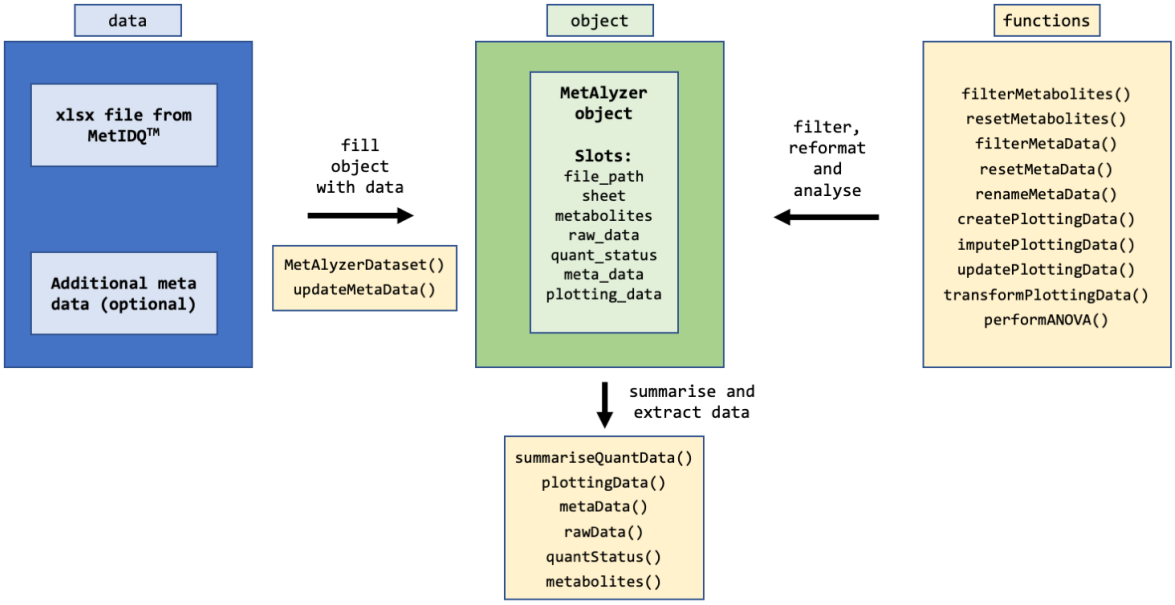

Figure 50: Overview of MetAlyzer functionalities. The S4 MetAlyzer object is filled with output data from MetIDQ™. The object can be filtered, reformatted, and analyzed. Additionally, functions to summarize and extract data are available. Colors indicate S4 object (green), data (blue), and R functions (yellow). This figure is also part of the vignette of the package which provides detailed information for users.

2.4.1 Comparison of extraction methods for intracellular metabolomics

The targeted metabolomics kit used for this study covers up to 630 metabolites from different chemical classes. These classes and the respective numbers of metabolites are displayed in Figure 51a. In an actual experiment, it may not be possible to quantify all metabolites simultaneously since reliable signal strength is required for effective quantification. A common approach is filtering by the LOD, defined as three times signal to noise ratio of the baseline. The filtered statistics for all extraction methods and sample types are shown in Figure 51b. The highest number of metabolites above the LOD was observed for liver tissue across all extraction methods (median: 391, range 171 - 456). The two investigated cell lines (HEK, median: 160.5, range 113 – 230 and HL60, median: 176, range 124 - 231) had very similar and intermediate total numbers of metabolites above LOD, while the lowest number was observed in bone marrow (median: 101, range 85 – 154).

The metabolite class with the highest number of quantified metabolites were triacylglycerols and glycerophospholipids which accounted for 52.7% of potentially covered metabolites. Overall, the distribution of metabolites above the LOD was similar to the distribution of metabolites in the kit. The best-performing extraction method was not identical for all sample types. *75 EtOH/MTBE B* and *100 IPA* yielded the highest and second highest coverages for both liver tissue and HL60. For HEK cells *75 EtOH/MTBE A* and *75 EtOH/MTBE B* performed best, while *100 IPA* and *75 EtOH/MTBE B* achieved the highest number of metabolites in bone marrow samples. In general, across the different sample types, the highest number of metabolites above LOD was detected in *75 EtOH/MTBE B* whereas *MeOH/ACN/H2O* achieved the lowest number. Notably, methods including handling with methanol provided on average a comparably low coverage (Figure 52a).

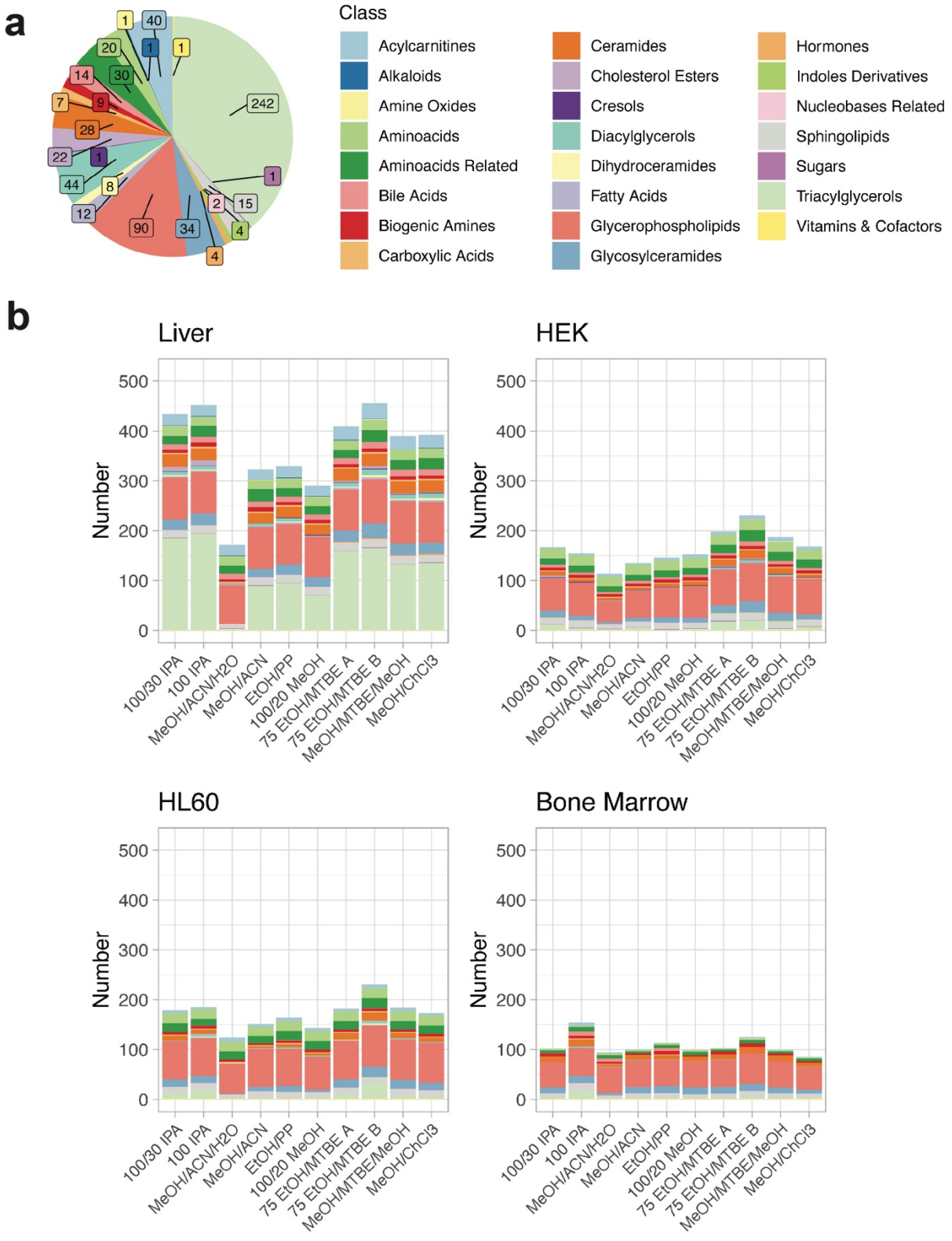

Figure 51: Metabolites in the kit and number of detectable metabolites. a) Metabolites covered by the Biocrates MxP® Quant 500 kit. Slices indicate the numbers of metabolites per class. b) Metabolites above LOD in the four different sample types for all extraction protocols. The color code at the top of this figure represents the different metabolite classes.

In addition to the number of metabolites above the LOD, it is crucial to take the repeatability of the extraction methods into account. To this end, I calculated the CV for replicates. Figure 52b shows the number of metabolites above LOD and filtered for a CV < 30% to penalize those protocols that show high variability between replicates. After filtering, the number of metabolites was reduced for all experimental conditions but showed a similar distribution. Liver samples provided the highest number of metabolites whereas bone marrow yielded the lowest number (Figure 52a,b). Focusing on the overall distribution of median CVs for the extraction methods and sample types, most combinations showed low variation between replicates (around 20%). However, some outliers were highly variable, such as bone marrow using *100 IPA* or *MeOH/ACN/H20*, HEK using *EtOH/PP,* and HL60 using extraction protocol *100/20 MeOH (*Figure 52c*).* Figure 52d displays the distributions of CVs as histograms for the metabolites detected above LOD. For all experimental settings, the distributions showed a long tail towards high CVs. Hence, a small number of metabolites present low analytical repeatability and it is not advisable to include them into the group of efficiently quantified metabolites.

Additionally, I calculated the sum of concentrations (SOC) i.e., the sum over all metabolite concentrations to estimate the overall analytical repeatability between biological replicates (Figure S 18). Depending on the extraction protocol and tissue, a consistent SOC over the biological replicates could be observed. For example, for *MeOH/ChCl3* the variability for liver and bone marrow was low (CV: 0.03 and 0.05), while HEK and HL60 displayed high variability (CV: 0.18 and 0.44). Due to the different sample types (liver tissue, bone marrow cells, and cell lines), the units for concentrations were given as picomole per mg and picomole per 10^6 cells, respectively. Hence, for liver tissue, the SOC is displayed on a different scale and the overall lower absolute number did not contradict the observation that liver tissues performed best based on the LOD as a quality control (QC) metric. For HEK, HL60, and bone marrow, an overall trend from high to low SOC for the sample types was observed, respectively. Comparing the SOC with the number of metabolites above the LOD indicated that these two measures did not correlate. Thus, the SOC provided additional information. For example, *100 IPA* performed best in most tissues based on the number of metabolites above the LOD while the SOC was comparable for multiple protocols including *EtOH/PP* and *75 EtOH/MTBE B* (Figure S 18 and cf. Figure 51).

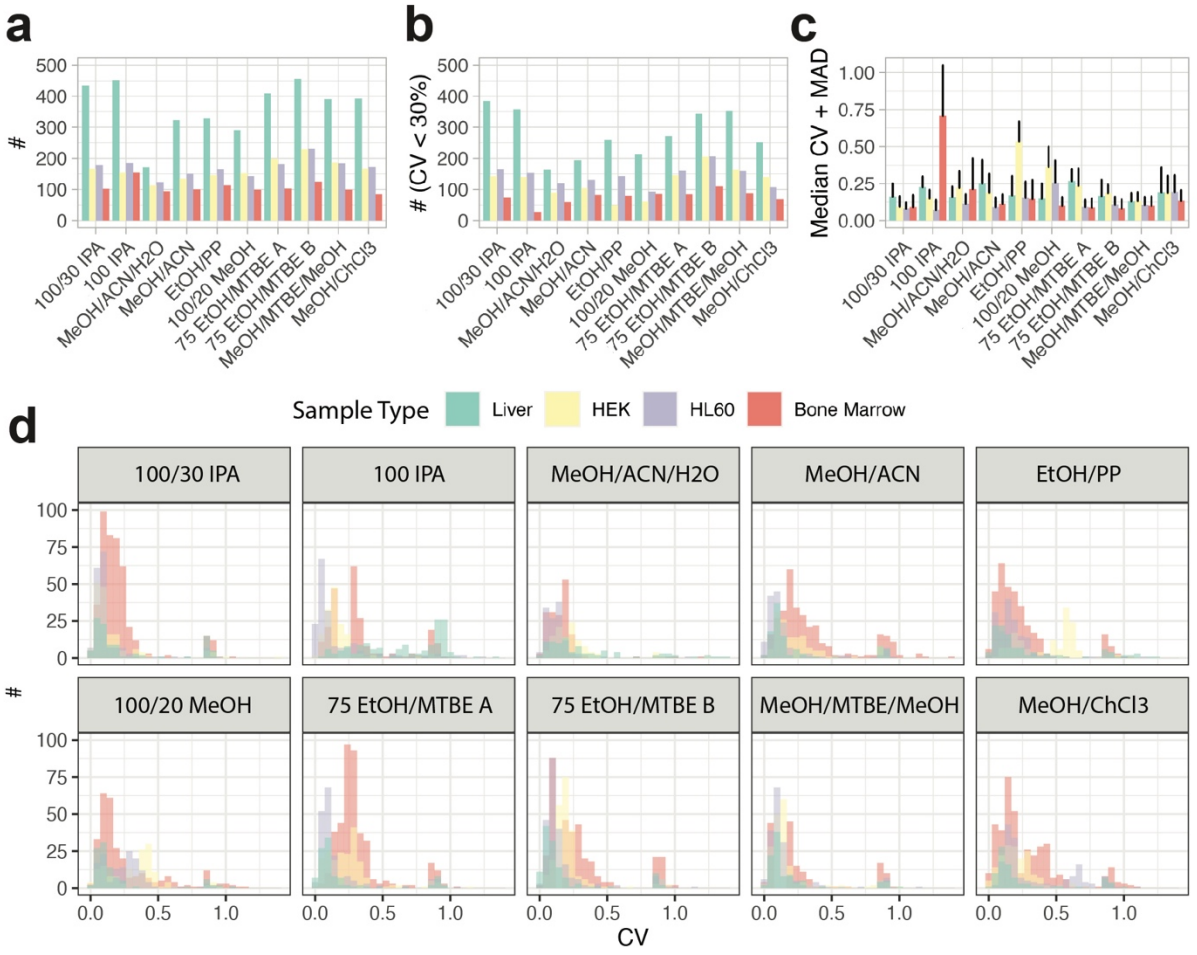

Figure 52: Statistics indicating repeatability for each extraction protocol and sample type. a) Bar plot showing number of metabolites above LOD. b) Bar plot showing number of metabolites above LOD and CV < 30%. c) Median CV and median absolute deviation (MAD). d) Histograms showing distributions of CVs for extraction protocols and sample types (only metabolites above LOD). Color code indicating sample type applies to all items.

In order to provide an unsupervised analysis of all study samples, I performed a PCA on all metabolite concentrations. This analysis revealed that the first five PCs were mostly associated with sample types as main effect, while PC3-6 and PC9-10 were associated with the extraction methods (Figure 53a,c). When clustering only liver tissue, similar protocols clustered together, and the main variability was determined by the solvents used in the extraction methods. PC1 captures about 50% of total variance and was strongly associated with the use of IPA (isopropanol), but also influenced by methanol (MeOH) and acetonitrile (ACN) (Figure 53b,d). Among the metabolites with highest loadings for PC1 were triacylglycerides, while ceramides and sphingomyelins seemed to drive the variability of PC2. The PCA showed that extraction efficiencies for metabolites of different classes with different chemical properties were highly influenced by the used solvents (Figure 53e).

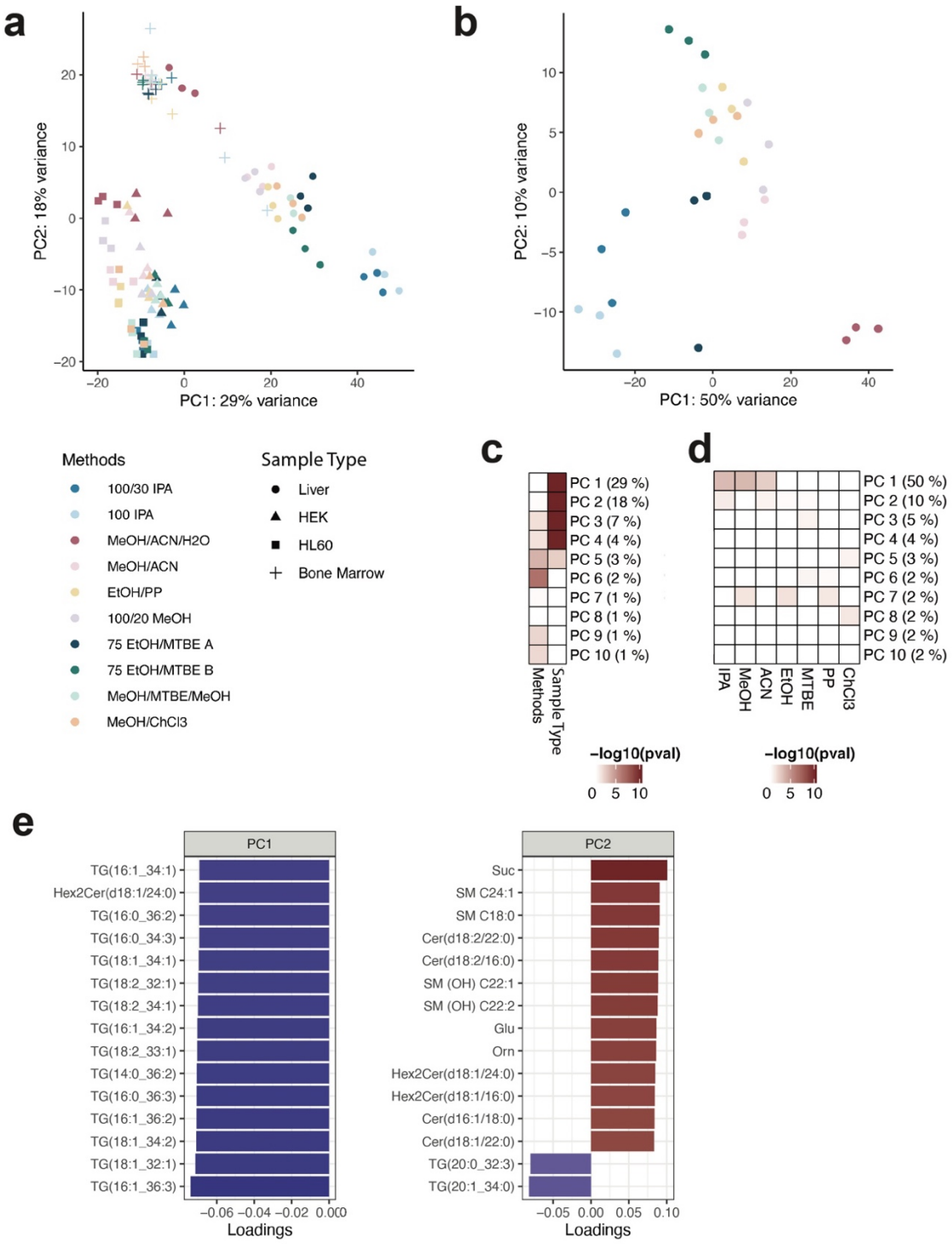

Figure 53: Clustering of extraction methods and sample types. a) PCA of different extraction methods across all sample types. b) PCA of different extraction methods for liver tissue. Colors and shape legend apply to both PCA plots. c) Heatmap showing the statistical association between sample type or extraction method and PCs. Color code represents p-values of Kruskal-Wallis Rank Sum Test. d) Heatmap showing the statistical association between solvents used in extraction methods and PCs. Color code represents p-values of Kruskal-Wallis Rank Sum Test. e) Bar plot showing the 15 highest loadings for PC1 and PC2 (ranked by absolute values). The color code represents signs of loadings.

2.4.2 Rationale for the choice of extraction protocol for SyTASC samples

Based on the work presented above, I aimed to identify the optimal extraction protocol for untargeted metabolomics in the SyTASC project, results of which were presented in subsection 2.3.7. I assumed that the sorted populations generated from PDX samples are most similar to the sample type “HL-60” in the comparative study, which is a confluent hematopoietic cell line. Taking only the number of metabolites above the LOD and the median CV into account, the best choice is the extraction protocol *75 EtOH/MTBE B*. This protocol resulted in the highest number of metabolites and a low median CV (Figure 54). However, the untargeted metabolomics approach is mainly designed for small molecules rather than complex lipidomics measurements. Therefore, I filtered out lipid or chemically lipid-like metabolite classes (acylcarnitines, ceramides, cholesterol esters, diacylglycerols, dihydroceramides, fatty acids, glycerophospholipids, glycosylceramides, sphingolipids, and triacylglycerols). Besides filtering for the metabolites above the LOD, I also excluded metabolites with a CV $\geq$ 0.3 and thus showed only low repeatability. Analogously to Figure 54, Figure 55 shows the statistics after these filtering steps. Filtering had a massive impact on the quality measures for the different extraction protocols. The total number of metabolites was largely reduced to about one quarter on average. For the *MeOH/ChCl3* protocol, there was no metabolite that fulfilled the quality requirement after filtering. As expected, the median CV was reduced for all extraction protocols. According to the new results, I chose the extraction method *MeOH/ACN/H2O* as a good trade-off between the number of metabolites and repeatability of the protocol represented by a low CV. This example emphasized the value of a cautious selection of pre-processing protocols, which can largely affect the scientific value of complex and time-consuming experiments such as metabolomic measurements.

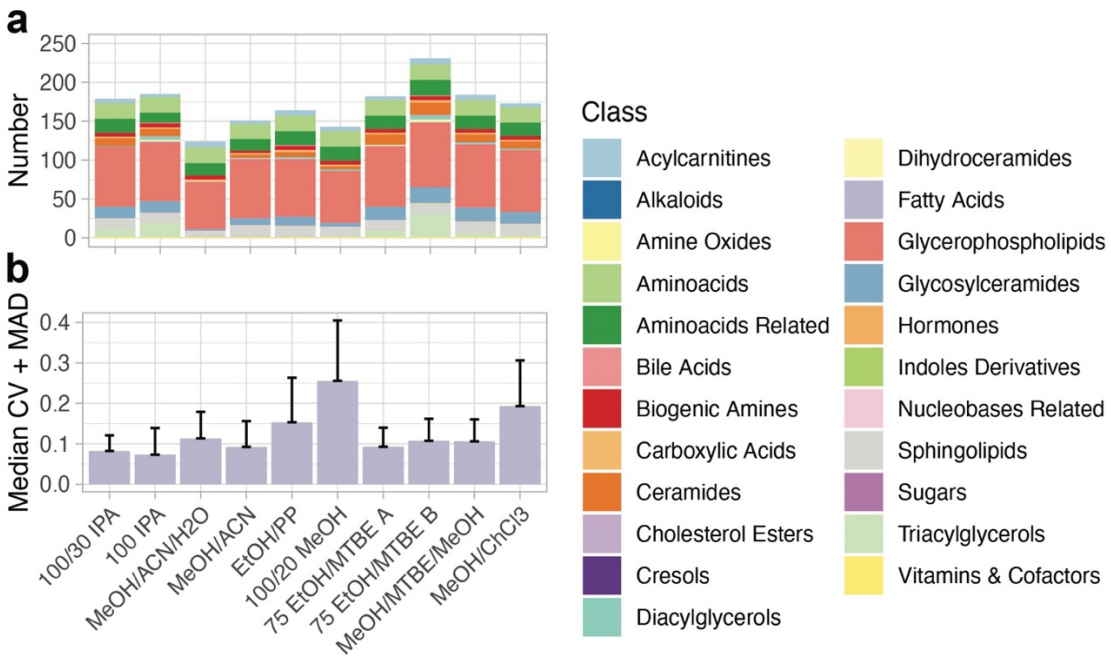

Figure 54: **Statistics across extraction protocols analogous to the available R Shiny app.** a) Stacked bar plot showing the number of metabolites quantified only by filtering metabolites below the LOD. b) Bar plot showing median CV including median absolute deviation (MAD) of metabolites above the LOD.

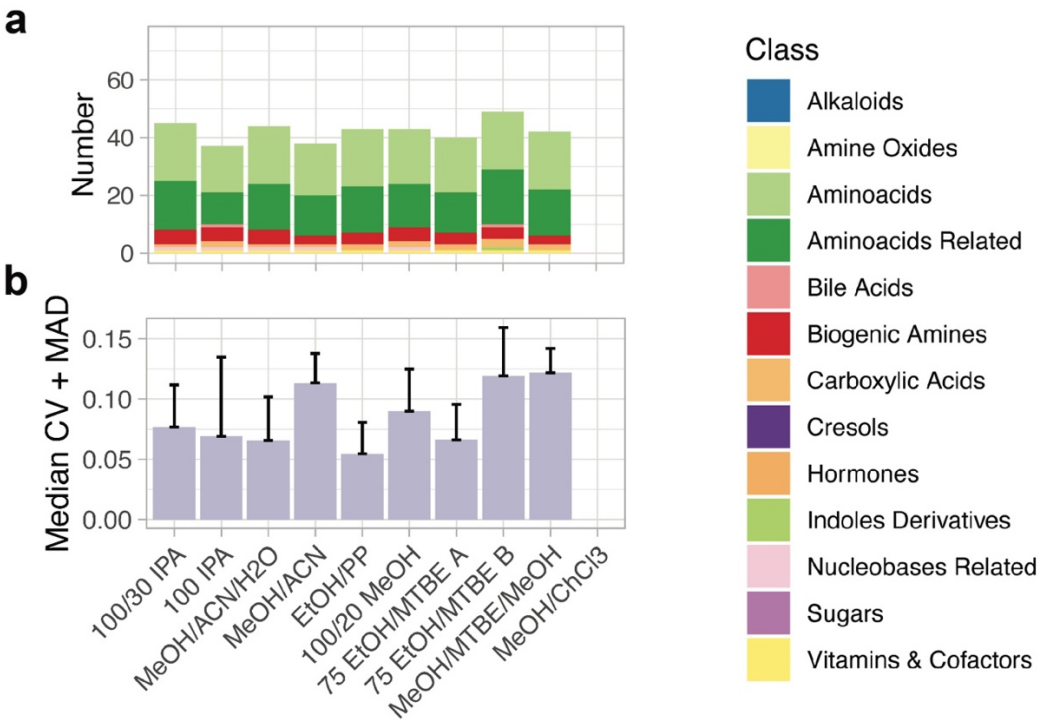

Figure 55: Statistics across extraction protocols based on filtering according to requirements of the SyTASC project. a) Stacked bar plot showing the number of metabolites quantified according to filtering options. b) Bar plot showing median CV and MAD calculated after filtering.

3 Discussion

The SyTASC project generated a multi-omics data set that allows comprehensive investigation of biological processes associated with early relapse of patients in a genetically homogeneous AML cohort by providing LSC-enriched cell populations. Analysis of this complex data set aimed to (i) characterize the five different populations derived from a novel FACS sorting strategy, including two engrafting populations, termed LSC-enriched, and (ii) investigate the differences between outcome groups (ER and LTR). I analyzed and integrated the available data layers (transcriptomic, epigenetic, and genetic information) in order to identify differentially affected biological pathways and train an outcome prediction signature that is also valid in external AML cohorts. Additionally, the outcome groups were characterized metabolically. The latter was based on (iii) a technical sub-project leading to the setup of an interactive platform for the identification of suitable extraction protocols for metabolomics studies.

3.1 A novel sorting strategy enriches functional and phenotypical LSCs

LSCs are of great interest in AML research since these cells produce all leukemic cells, and therefore their complete therapeutic eradiation is crucial for a lasting cure.[6] However, LSCs are often therapy-resistant, and their ability to self-renew and reinitiate the disease drives the relapse of patients.[9,10] Dr. Nadia Correia established a novel FACS sorting strategy to enrich these cells. In xenotransplantation assays using NSG mice she could show that only populations positively selected for GPR56 and negatively selected for NKG2DL engraft, independent of the CD34 immunophenotype. In total, five different populations were sorted and submitted to RNA-seq and methylation profiling.

These sorted populations contained different cell types, which accounted for the main variability in the data set. A strong stem-like phenotype of the engrafting LSC-enriched populations was observed. Embedding the samples into healthy single-cell data allowed to approximately map the differentiation stages of the sorted populations, which in a simplified model may be placed in a differentiation gradient (Figure 12). Figure 56 shows an overview of this model.

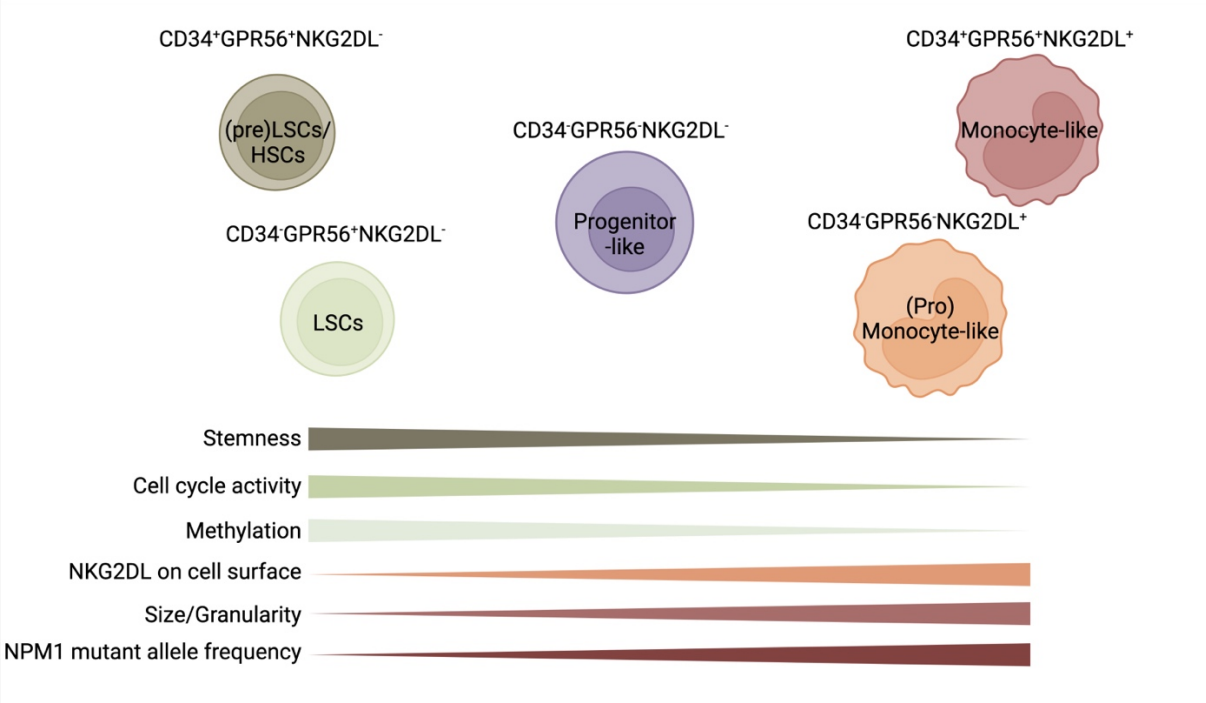

Figure 56: Overview of population characteristics.

3.1.1 Characteristics of the sorted populations

Hierarchic organization of leukemic cells has been observed in many AML samples.[40–42] In the SyTASC cohort, this was reflected by the similarity of the sorted AML cell populations to different cell types in healthy hematopoiesis. Engrafting populations had a more stem-like phenotype as illustrated by increased levels of the LSC17 score, LSC 104-genes correlation, and GSEA results.[51] Accordingly, the morphology was more stem-like; in the literature stem cells have been described as smaller in diameter and less granular since functional compartments are not yet developed.[98,178] AML is characterized by the accumulation of myeloid progenitors that cannot differentiate into functional mature blood cells.[41] Upon embedding into healthy hematopoiesis, these presumably not-functional leukemic blasts clustered to the (pro)monocytic populations, indicating a progression along a differentiation trajectory. Hierarchy within leukemic cells was furthermore supported by an assessment of mutant allele frequency. The number of mutant *NPM1* alleles was higher in the more differentiated cells indicating that these are the progeny of the dominant, malignant LSC clones (Figure 13).[32]

Analyses of the cell cycle showed higher proliferation activity for the engrafting than for the non-engrafting populations, which fits the hypothesis of a proliferating LSC and/or intermediate progenitor population that produces leukemic blasts. These blasts are not further proliferating and therefore are comparably less cycling. Interestingly, it has been described that overall AML cells do not proliferate more than normal hematopoietic cells, and therapy-persistent LSCs are commonly characterized by cell cycle quiescence.[41] With respect to proliferation, engrafting cell populations thus represent a heterogeneous mix of cells, but, as discussed above, are enriched for LSCs. Historically, LSCs were also referred to as leukemia-initiating cells (LICs) and were estimated to be a very rare population (1 in 1 million leukemic blasts).[49] The low abundance of these cells was reflected by the xenotransplantation results. Engraftment was almost exclusively observed in $CD34^{-}GPR56^{+}NKG2DL^{-}$ and $CD34^{+}GPR56^{+}NKG2DL^{-}$ populations, however, not all replicates showed engraftment (Table S 1). Taken together, this indicates enrichment of slowly cycling therapy-resistant LSCs at the apex of AML and more progenitor-like cells driving the replenishment of leukemic blasts. Bioinformatic estimation of the cell cycle in bulk RNA-seq data only reflects averaged cell cycle fractions but does not allow analysis of differences between heterogenous subpopulations. Engraftment of non-sorted AML samples in xenotransplantation assays has been reported for many AMLs.[195] Considering that engraftment occurs even at very high dilution of cells that actually mediate engraftment, the absence of engraftment and thus the depletion of LSCs in the more differentiated populations might even further support the enrichment of LSCs in the $CD34^{-}GPR56^{+}NKG2DL^{-}$ and $CD34^{+}GPR56^{+}NKG2DL^{-}$ populations.

Along the hierarchical gradient illustrated in Figure 56, increasing hypomethylation was observed. This is consistent with publications describing a general hypomethylation along myeloid differentiation, although dynamical patterns at different stages were observed .[108,109] In contrast, DNA methylation increases with differentiation in other tissues, including lymphoid lineage.[107] A former study by Jung et al. investigating DNA methylation in FACS-purified populations showed that differentially methylated regions in LSC-enriched/engrafting populations are mostly hypomethylated compared to non-engrafting populations.[196] For LSC enrichment, Jung et al. used CD34 and CD38 as FACS markers. However, as opposed to our setting, only 3 out of 15 patients in that cohort harbored a *DNMT3A*-R882H mutation. It is thus very likely that the mutation status of *DNMT3A* plays a crucial role in the methylation pattern.

Methylation patterns are highly specific for cell populations in healthy hematopoiesis.[108] Interestingly, in the SyTASC data, the same pattern of decreasing methylation with increasing differentiation was also observed for healthy samples when considering the $CD34^{+}GPR56^{+}NKG2DL^{-}$ population the most HSC-enriched population in the healthy setting. While, as described by Ji et al., methylation decreases with myeloid differentiation, Hodges et al. observed that based on methylation patterns, HSCs are still more similar to differentiated myeloid cells than to lymphoid cells. [108,197] When investigating the cell of origin in the SyTASC data by clustering of methylation data, I showed that all sorted populations were in the same clade as GMPs, except for one $CD34^{+}GPR56^{+}NKG2DL^{-}$ sample that clustered to LMPPs (lymphoid-primed multipotential progenitors) (Figure S 4). As suggested by Goardon et al., LSCs seem to arise from progenitors that acquired self-renewal properties rather than from HSCs and most closely resemble GMPs.[198] However, it must be considered that the reference data set only comprised HSPCs. Therefore, a more comprehensive reference might have shown more subtle effects in the dendrogram. Still, the methylation clustering recapitulated the findings of a study that proposed the coexistence of LMPP-like and GMP-like LSCs in AML.[198] These observations indicate that GMPs may be the cell of origin of the disease here.

An observation in the exploratory analyses was the high expression of immunoglobulin genes in LSC-enriched populations. Most likely, this was due to ambient mRNA attached to the surface of the cells.[179] This interpretation was supported by the lack of intronic mRNA. Hence, I removed immunoglobulin genes from further analyses. However, the difference in abundance between the populations was striking, which could either indicate that LSC populations reside in a specific niche or have different surface properties that facilitate the attachment of these mRNAs. Even though the function or origin of immunoglobulin mRNA could not be determined, it is yet another marked difference between on one hand the LSC-enriched and on the other hand the more differentiated populations.

Immune checkpoint therapy has largely improved the treatment of many solid tumors.[199] For AML, these therapies showed less promising results.[175,200] Among the most prominent differences between engrafting and non-engrafting populations in the SyTASC data were genes and gene sets related to immunoregulatory processes. Most likely, these expression patterns represent the differentiation stages of the populations. [201] The expression pattern of immune checkpoints could be one of the reasons for failing immune therapy in AML. Many

classic inhibitory immune checkpoint genes were not expressed in the LSC-enriched populations. One example was the apparent lack of expression of LILRs. LILRB4 has been described as a promising target for monocytic AMLs.[201] It's lack of expression in engrafting LSCs raises the question if a sustained response can be expected or if targeting this receptor would only affect the leukemic blasts. Taken together, this work gives hints to why checkpoint inhibition does not target LSCs and hence fails to eradicate the most critical AML cells.

3.1.2 The "NK-depleted" sorting strategy reveals differences in LSC-enriched populations

The analysis of immune checkpoint genes also showed high levels of KIRs in the $CD34^-GPR56^+NKG2DL^-$ populations. These genes are specific for NK cells.[202] This was in line with the observation of $CD34^-GPR56^+NKG2DL^-$ clustering to NK cells when embedding the samples in normal hematopoiesis. The varying abundance of remaining NK cells was problematic for purity of the extracted expression signals from the LSC-enriched populations. These remaining NK cells were present in the $CD34^-GPR56^+NKG2DL^-$ but not in the $CD34^+GPR56^+NKG2DL^-$ population, leading to a great bias in the differential expression analysis. The FACS marker CD34 enriches HSPCs and therefore excludes differentiated immune cells such as NK cells in $CD34^+$ populations. Additionally, GPR56 was described as a marker for NK cells. Thus, positive selection for this marker might facilitate the enrichment of NK cells in the $CD34^-GPR56^+$ population.[97] In the "original" data set, a few samples were CD3-depleted, which accounts for T cells and macrophages but not NK cells. A subset of samples was re-sorted by Dr. Elisa Donato using an improved FACS sorting strategy referred to as "NK-depleted". Besides CD3, the "NK-depleted" sorting strategy also comprised CD19, CD235a, and CD20 to exclude lineage-positive cells in the $CD34^-$ population. For the "NK-depleted" data set, the estimated NK cells via deconvolution were marginal and allowed effective comparison between the two engrafting populations ($CD34^-GPR56^+NKG2DL^-$ and $CD34^+GPR56^+NKG2DL^-$) (Figure 18). All other analyses were not affected by the presence of NK cells. In particular, in all comparative analyses of the two outcomes groups (ER and LTR) shown in section 3.2, confounding effects due to remaining NK cells can most likely be excluded as both groups showed comparable abundances of NK cells. Additionally, one might speculate that the improved, "NK-depleted" sorting strategy with much lower remaining NK cells in the CD34- fraction leads to an even higher enrichment for LSCs, the cells of interest.

Overall, based on the clustering of transcriptomic data and on the low number of significantly differentially expressed genes, the $CD34^+$ and $CD34^-$ engrafting populations appeared to be very similar. However, the direct comparison revealed differential activity of important transcription factors, e.g., high activity of SOX2 in the $CD34^+$ LSC-enriched population. The role of this transcription factor in stem cell development and maintenance of stemness has been extensively studied in various stem cell populations and tissues.[181] Interestingly, Sox2 interacts with Npm1 in mice, which is mutated in all samples of the SyTASC cohort.[203] However, a role of *SOX2* in hematopoiesis has not been described in literature.[204] In the SyTASC data set, strong enrichment for JAK-STAT signaling in the $CD34^+$ engrafting population was observed, and overexpression of *SOX2* has been described to be induced by JAK2/STAT3 activity. Signals from different sources of activation may thus be convoluted.[205] The JAK-STAT pathway is a tightly regulated signaling cascade of Janus kinases (JAK) and signal transducers and activators of transcription (STATs), which regulates hematopoiesis, HSC proliferation, survival, and self-renewal.[165] Constitutive activation of this pathway has been widely described for hematologic malignancies, including AML, even leading to inhibitors of JAK being approved for use in patients.[206] A recent publication showed enrichment of JAK-STAT signaling and promotor accessibility in normal HSCs compared to *DNMT3A*-mutant HSCs indicating a possible deviant activity in these mutant HSC populations.[207] This is in line with higher LSC17 scores, higher abundance of normal HSCs and enrichment of HSC-related gene sets in $CD34^+$ population when compared to the $CD34^-$ population (Figure 19 and Figure 21). Strikingly, some of the $CD34^+$ samples which had particularly low *DNMT3A*- and *NPM1*-mutant allele frequencies showed multi-lineage engraftment in xenotransplantation assays. Taken together, these were strong signs of the presence of retained healthy HSCs in the $CD34^+$ population. This is consistent with common usage of CD34 to enrich HSPCs, and negative selection for CD34 excludes normal HSCs.

3.1.3 Enriched populations from healthy bone marrow are different from AML

Isolation of hematopoietic populations by FACS-sorting has been intensively studied but markers used within healthy hematopoiesis cannot necessarily be transferred to the diseased system and vice versa.[93] In order to elucidate this further, in addition to the AML samples, the sorting strategy was also applied to healthy bone marrow samples. Overall variability between the sorted populations was higher than between the different disease statuses. For the most differentiated populations $CD34^-GPR56^-NKG2DL^+$ and $CD34^+GPR56^+NKG2DL^+$, PCA

on transcriptomic data showed intermingled clustering of healthy and AML samples. This was in line with the embedding of healthy samples into single-cell data from healthy hematopoiesis revealing that these populations, like AML, clustered to monocytes (Figure S 3 and Figure 12). One might speculate that using a reference data set more fine-grained in the differentiated populations would show more distinct clustering. In contrast, PCA clustering of methylation data showed clear differences between healthy and AML $CD34^{-}GPR56^{-}NKG2DL^{+}$ and $CD34^{+}GPR56^{+}NKG2DL^{+}$ populations (Figure 8).

The healthy counterpart to the AML progenitor-like population $CD34^{-}GPR56^{-}NKG2DL^{-}$ also clustered relatively diffuse when embedded in the healthy reference samples. The counterparts to the two LSC-enriched populations, $CD34^{-/+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$, showed a different picture. While leukemic cells clustered mostly to HSCs when embedded into normal hematopoiesis, healthy populations clustered to erythroid progenitors. Accordingly, clustering to a healthy methylation reference showed that the healthy $CD34^{-/+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$ populations were most similar to CMPs (common myeloid progenitors) or MEPs (megakaryocytic-erythroid progenitors), which are progenitors that also give rise to erythroid cells.[208] However, when clustering AML samples to a healthy methylation reference, all samples showed the highest similarity with GMPs. Hence, the $CD34^{-/+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$ populations from on one hand AML and on the other hand healthy samples behaved differently based on transcription and methylation data. As described above, stemness signatures and the abundance of HSCs were different for AML and healthy $CD34^{-/+}GPR56^{+}NKG2DL^{-}$ and $CD34^{-}GPR56^{+}NKG2DL^{-}$ populations. Consequently, healthy and AML populations are not directly comparable since they do not recapitulate the same types of cells. Hence, a direct comparison between healthy and AML populations was not informative for most questions and the analyses of the populations suggested that sorting strategies might not be transferrable from healthy hematopoiesis to the diseased system.

3.1.4 Advantage of combining the different markers to enrich for LSCs

NKG2DL status explained the largest fraction of the variance observed in RNA-seq and methylation data in various complementary analyses. In line with observations by Paczulla et al., the NKG2DL immunophenotype was strongly associated with the differentiation stage of the sorted populations.[98] However, only those cells which were additionally selected for

GPR56 positivity engrafted in xenotransplantation assays. Even though both markers alone have been shown to enrich for LSCs in AML, the complementary populations, $CD34^-GPR56^-NKG2DL^-$ and $CD34^+GPR56^+NKG2DL^+$ did not engraft.[94,98] Therefore, a combination of both markers most likely enriches for LSCs in with greater purity. Based on morphological markers and on clustering of the data, the $CD34^-GPR56^-NKG2DL^-$ population was phenotypically similar to the engrafting LSC-enriched populations ($CD34^{-+}GPR56^+NKG2DL^-$ and $CD34^-GPR56^+NKG2DL^-$). However, in other analyses such as embedding into normal hematopoiesis it behaved quite differently. Direct comparison between the $CD34^-GPR56^-NKG2DL^-$ and the engrafting LSC-enriched populations revealed processes involved in cell cycle and differentiation as major differences. This was also supported by a lower LSC17 score in the $CD34^-GPR56^-NKG2DL^-$ population. Hence, this population seemed more differentiated and therefore lost its potential for engraftment, while immunophenotypically losing GPR56 positivity. However, a specific pathway driving this development could not be identified. Additionally, it could not finally be determined if the loss of GPR56 positivity is the reason for or a consequence of the loss of stemness. Since the $CD34^+GPR56^+NKG2DL^+$ population showed a differentiated phenotype with no engrafting potential, the latter is more likely. Of note, the immunophenotypic absence of NKG2DL showed a higher impact on engraftment potential than the presence of GPR56. However, as stated above, absence of NKG2DL alone seemed to be a necessary but not sufficient marker for the enrichment of engrafting leukemic stem cells in this sorting scheme. In summary, the combination of GRP56 and NKG2DL enriches functional and phenotypical LSCs, and the combination showed an added value for a purer enrichment of LSCs.

CD34, together with CD38, is one of the most commonly used markers in sorting strategies to enrich for healthy HSC, and is also intensively used to study LSCs.[42,94] However, publications by Pabst et al. and Paczulla et al. already showed that GPR56 and NKG2DL are markers for LSCs independent of CD34.[94,98] The SyTASC data confirmed these findings and even revealed that positive selection for CD34 and GPR56 does not necessarily enrich for LSCs since the $CD34^+GPR56^+NKG2DL^+$ did not show engraftment (except for one sample; Table S 1). This independence from CD34 is of particular interest when studying CD34-negative AMLs, which account for about 25% of all cases. Especially, *NPM1*-mutated AMLs are enriched for samples that lack expression for CD34, defined as CD34 present on <10% of all leukemic blasts.[41] In those samples, most LSCs seem to reside in the $CD34^-$ fraction, whereas only a few showed a $CD34^+$ immunophenotype.[93,101] Interestingly, Quek et al. hypothesized that based on

transcription, $CD34^+$ and $CD34^-$ LSCs represent the same cell with different immunophenotypes and already advocated for alternative markers for LSC enrichment.[209] This is clearly in line with findings in this study, which showed high similarity between these populations and minor transcriptional differences that most likely originated from retained normal HSCs in the $CD34^+$ population as described above. The novel sorting strategy overcomes the challenge of enriching for LSCs in CD34-negative AML, as shown for the *NPM1*-mutated SyTASC cohort presented here. The novel sorting strategy and insights gained are presented in a manuscript currently in revision at *Blood Advances* (cf. chapter "Own publications"). In summary, multi-omics analyses shed light on the characteristics of the different sorted populations and were in accordance with the experimental findings on engraftment potential by xenotransplantation assays.

3.2 Two distinct outcome groups in a genetically homogenous cohort

Understanding the biological changes associated with relapse of AML patients is crucial to improve treatment. This includes identifying genes or biological pathways that could be used as a therapeutic target and for stratification of patients. Particularly since cancer therapies are associated with major, sometimes life-threatening, side effects and quality-of-life impairments for patients, selecting an effective therapy option is of utmost importance.[210] The following subsections discuss the findings and differences observed between the two outcome groups (ER and LTR) in the SyTASC cohort.

3.2.1 A genetically homogenous cohort to study LSCs

The SyTASC cohort was retrospectively selected for *NPM1*-W288fs*12- and *DNMT3A*-R882-mutant patients. However, other mutations co-occurred in a subgroup of patients, e.g., a large number of patients also carried an *FLT3*-ITD mutation which showed a low but statistically significant association with the outcome groups (Table 3). Nevertheless, this could not sufficiently explain the dramatic RFS difference between the two outcome groups. Another possible reason for the different RFS could be a non-dominant subclone that re-initiates the AML after chemotherapy. This is unlikely since the outcome group is, particularly in the LSC-enriched populations, a major source of variability in RNA-seq and DNA methylation data which cannot be caused by a low abundant subclone (Figure 28 and Figure S 10). Particularly the integration using MOFA showed an association of the outcome groups with different LFs reflecting high variance explained, which are unlikely due to clones of minor frequency (Figure 7 and Figure 31).

All samples were collected at diagnosis, went into complete remission after chemotherapy, and were retrospectively stratified by the time until relapse. Available clinical information was analyzed to identify potential confounding factors. For example, age is a major determinant of survival in AML and cancer in general.[211] The SyTASC cohort includes only samples from young adults aged 22 to 65 years, and age showed no significant association with the outcome group. Among the tested features, only BM blast count showed a low and statistically significant increase in the hazard ratio (Table 4). High levels of BM blasts have been described as an indicator of poor prognosis, potentially due to a more advanced disease progression.[189] Notably, the percentage of BM blasts is also a diagnostic parameter. BM samples with more

than 20% blasts are considered leukemic.[212] The percentage of BM blasts for the SyTASC cohort was reported between 4% and 93%, which is not in line with the requirements for diagnosis. Hence, it would be questionable if this data is valid to be used as a confounder. Since samples originate from two hospitals (Dresden and Ulm), a bias in acquiring this parameter is also likely, potentially due to personnel, equipment, or different standard operating procedures (SOPs), which are frequent sources of bias in clinical information.[213] In summary, the SyTASC cohort showed a homogeneous genetic background and no major confounding factors could be identified. However, the batch effect from RNA-seq data remained, which had to be considered in all analyses comparing the outcome groups (cf. Figure S 1).

It is essential to target LSCs for sustained therapy success.[41] Hence, for the analysis comparing the outcome groups, I focused mainly on the LSC-enriched populations. This approach was also supported by the more pronounced differences in these two populations compared to analyses including the more differentiated ones (Figure 27, Figure 28 and Figure S 10). For example, unsupervised PCA of RNA-seq and DNA methylation data showed that the first PCs were statistically significantly associated with the outcome group in the LSC-enriched populations. Interestingly, the association with the outcome group as the major driver for variability was even stronger than the patient-specific effect. Still, patient-specific differences are a frequent source of variability also in the SyTASC data set. For example, many of the LFs identified by MOFA with lower explained variance were specific for individual patients (Figure 7).

In contrast, analysis of all cell populations merged showed less variability between the outcome groups. Hence, one might speculate that the leukemic phenotype is more similar in the differentiated populations. A similar observation was made for the RNA-seq bulk data from unsorted samples. Comparison between outcome groups in these samples revealed only 40 significantly differentially expressed genes. However, some of the significantly differential processes discovered in LSC populations (e.g., MHC-II overexpression in ER samples) were also different in bulk samples even though significance was lacking. In summary, the differences between the outcome groups were more pronounced in the LSC-enriched population. This may support the hypothesis that the crucial differences leading to differential RFS are rooted in the LSCs and that these cells are indeed responsible for relapse.

3.2.2 ER samples present a more stem-like phenotype

A critical mechanism for the resistance to anti-proliferative chemotherapy in LSCs is likely a quiescent or dormant state allowing them to sustain therapeutic interventions.[30] Dormant HSCs, and most likely their malignant counterparts, are at the apex of the respective hierarchies.[214] Healthy dormant HSCs have been characterized by their extremely low proliferation rates.[28] Immunophenotypic analyses considered about 20-30% of all HSCs as dormant.[29,214] As dormancy is intimately linked to stemness, differences in stemness between the outcome groups were investigated. Analysis of LSC17 scores and correlation with LSC 104-genes signature as well as GSEA showed higher enrichment for stemness in ER compared to LTR samples (Figure 29). Additionally, the distance to healthy HSCs was smaller for ER samples than for LTR samples when embedded into healthy hematopoiesis (Figure 30). This does not necessarily prove a more dormant phenotype but indicates a more immature phenotype of the ER compared to the LTR LSC-enriched populations. However, the inferred cell of origin by comparison to HSPC methylation data indicated a similarity to GMPs but no difference between the outcome groups (Figure S 4). Consequently, the more stem-like transcriptional phenotype of ER samples could be either due to a priori differences during the onset of AML or posterior during to progression of the disease. Assuming that the cell of origin is indeed GMP-like for both outcome groups, the ER phenotype might be more accurately described as dedifferentiated. Analysis of the cell cycle phases did not reveal any difference between the outcome groups (data not shown). Even though the SyTASC sorting strategy strongly enriches for LSCs, likely few cells in the sorted populations show a dormant, chemo-resistant phenotype. Hence, specific patterns and differences of these low abundant cells may be diluted in the “population bulks”. This is particularly likely when considering the observation of overall high cell cycle activity in the LSC population discussed in subsection 3.1.1.

The complex and context-dependent regulation of normal hematopoiesis by TGFβ signaling has been intensively studied.[190,215] TGFβ signaling has been described as crucial for stem cell quiescence and the related BMPs are essential regulators in the HSC microenvironment maintaining hematopoietic progenitors in an undifferentiated state.[215,216] In the SyTASC data set, significantly higher TGFβ signaling activity was observed in LTR samples compared to ER samples by pathway enrichment analysis and was supported by striking enrichment of related genes and gene sets (Figure 39). This is not in line with the hypothesis that ER samples might be more quiescent than LTR samples. In contrast, other studies observed BMPs as initiators

of differentiation.[217] Particularly in the leukemic context, the role of TGFβ and BMP signaling was ambiguous in different studies depending of the AML subtype.[46,190,218,219] Recently, Sun et al. showed that inhibition of BMP signaling promotes self-renewal in acute myeloid leukemia cells and thus BMP signaling may act as a tumor suppressive pathway.[220] These findings are in line with the higher TGFβ and BMP signaling activity in LTR samples which were characterized as more differentiated compared to ER samples. Interestingly, the striking upregulation of the BMP signaling inhibitor noggin (NOG) in LTR samples does not fit the downstream activation of BMP signaling.[221] A potential explanation might be a posttranslational regulation of noggin. Additionally, the enrichment of BMP downstream signaling might be driven by the strong activation of the TGFβ pathway which may outcompete inhibition by upregulated noggin. To investigate how stemness signatures and TGFβ signaling are regulated in the LSC environment, further research is needed.

Expression of MHC-II genes has already been described in the 80s for some AMLs.[222,223] Later studies showed that downregulation and loss of surface presentation are associated with relapse in AML patients.[177,224] The strongest differential signal between the outcome groups was the upregulation of MHC-II transcription in ER samples. This observation was present in all sorted populations, and the bulk RNA-seq showed clear differential expression. Hernandez-Malmierca et al. showed that HSPCs constitutively present antigens via MHC-II as an immunosurveillance mechanism.[225] High expression and surface presentation of MHC-II have also been associated with stemness in AML patient samples and cell lines. Interestingly, the expression levels also seem to be dependent on the genetic status of the AML (e.g., *FLT3*-ITD).[225] However, in the LSC populations of the SyTASC cohort, neither a specific association with *FLT3*-ITD status nor differential expression of relevant immune checkpoints could be observed (data not shown). In summary, upregulation of MHC-II transcription fits to the observation of enhanced stemness in ER samples.

3.2.3 Alteration of energy metabolism in engrafting LSC populations

Reprogramming of energy metabolism in cancer cells has been intensively studied and is considered a hallmark of cancer.[167] Accordingly, rapidly proliferating cancer cells rely on glycolysis rather than oxidative phosphorylation as the primary mode of energy production.[167] Comparison of the outcome groups in the SyTASC data showed that LSC-enriched populations sorted from LTR samples seemed to utilize glycolysis for energy production. In contrast, LSC-

enriched ER samples showed evidence for higher oxidative phosphorylation activity. Interestingly, the analysis of bulk RNA-seq data also displayed enrichment for oxidative phosphorylation and mitochondrial complexes in ER samples but statistical significance was lacking. For HSCs and LSCs, it has been hypothesized that these cells generally favor glycolysis over oxidative phosphorylation, particularly since they reside in a relatively hypoxic niche in the bone marrow.[161,226,227] However, recent studies have shown that especially quiescent LSCs or resistant AMLs can be targeted via inhibition of oxidative phosphorylation.[86,162,228–230] These studies are in line with the observation of higher oxidative phosphorylation dependency in ER samples with a resistant, potentially more quiescent phenotype. In contrast, LSCs in LTR samples presented a chemo-sensitive phenotype and exhibited a more glycolytic metabolic state, potentially resulting from a more differentiated cell state. Interestingly these effects were more dominant in the $CD34^{+}GPR56^{+}NKG2DL^{-}$ compared to the $CD34^{-}GPR56^{+}NKG2DL^{-}$ population. This may be caused by the presence of retained normal HSCs, which are potentially more abundant in LTR samples based on the observation of multi-lineage engraftment of AML bulk samples (Table S 1, cf. subsection 3.1.2). Hence, these more glycolytic, healthy HSCs potentially contribute to a convolved signal of observed difference in energy metabolism.

In addition to transcriptomic, epigenetic and mutational profiling, a subset of samples derived from the SyTASC cohort were subjected to metabolomics, however using a slightly different sorting strategy for enrichment of cell types. Metabolomics analysis revealed differences between the outcome groups which were in line with observations in the corresponding populations by RNA-seq (Figure 37). However, statistical evidence was low and only few metabolites showed significantly differential abundance. An explanation might be the fast turnover of metabolites, particularly those involved in energy metabolism.[154] Pre-analytical sample processing for metabolic measurements has been described to critically affect its results.[155] Preparation of the samples from the SyTASC cohort included FACS-sorting, which may not be suitable for measuring differences in energy metabolism. This procedure contains multiple time-consuming steps, including staining. In contrast, sampling for metabolomics usually recommends direct cooling of cells or tissues[231–233]. Particularly for the $CD34^{+}GPR56^{+}$ population, the number of significantly differentially abundant metabolites was relatively low with only four metabolites. In contrast, the most significantly enriched metabolic pathway was PI3K signaling in ER samples, also reflected by the differential abundance of inositol bisphosphate (hexose bisP). This was in line with the observed higher activity of PI3K signaling

in transcriptomic data of ER LSC-enriched samples. Constitutive activation of PI3K/AKT/mTOR signaling is often hyperactivated in AML and associated with poor survival.[158] A recent study investigated the gain of accessibility of promotors in HSCs with induced *Npm1* and *Dnmt3a* mutations and showed enrichment for PI3K/AKT/mTOR signaling together with mitochondria- and stem cell-related pathways.[207] These findings are overall in line with the above described observations for the LSC-enriched population in ER samples. Metabolomics is a complex data type and conclusions from the metabolomics data of the SyTASC samples have to be drawn with care, but further mining of the already generated data may still yield additional insights in the future.

3.2.4 Mutant allele frequencies and timing of mutations

AML has been described to comprise a mixture of genetically distinct subclones. One method to investigate clonal relationships is to infer the allele frequency of mutated genes.[76] In the DNA derived from the SyTASC samples, *DNMT3A* and *NPM1* variants were determined by panel sequencing or WGS. For the sorted cell populations deeply analyzed in this thesis, I computed the mutant allele frequencies based on mutated RNA-seq read counts using the prior knowledge of the expected variant positions. Hence, I termed this analysis "mutant allele frequency" in contrast to variant allele frequencies (VAF); a term reserved for the analysis of DNA sequencing data. Overall, the mutant allele frequency of *DNMT3A* was higher than that of *NPM1*. *DNMT3A* has been described as a pre-leukemic mutation priming the cell, and subsequent mutations such as *NPM1* then initiate AML.[32] In most samples, the *DNMT3A* mutant allele frequency was around 50%, confirming that on average most cells harbor this heterozygous pre-leukemic mutation. In contrast, about 40% of *NPM1* alleles were mutated (Figure 13). Thus, potentially not all sorted cells were leukemic. Interestingly, the mutant allele frequency was significantly higher in the ER samples compared to LTR, indicating either lower abundance of non-leukemic cells or a potentially more progressed AML with a dominant *NPM1*-mutated clone. For *DNMT3A*, the mutant allele frequency was not significantly altered between the outcome groups. Particularly the *DNMT3A*-R882H mutation, which is the main variant in the SyTASC cohort (besides R882C and R882G), has been described to cause a loss of methyltransferase activity and, therefore, hypomethylation.[62] Despite absence of differences in mutant allele frequency, global analysis of methylation differences showed an increased hypomethylation in ER compared to LTR samples (Figure 32 and Figure S 11). One explanation could be a different cell of origin since methylation is highly

specific for cellular differentiation stages.[108,234] However, as described above, when samples were clustered to methylation patterns of normal HSPCs, a difference between the outcome groups could not be observed (Figure S 4). Additionally, the ER samples showed global hypomethylation, which would not fit a more stem-like phenotype suggested by transcriptomic analyses. These observations rather argue against a different cell of origin between ER and LTR. In general, when compared to healthy BM samples, AML samples showed clear global hypomethylation, also in the more differentiated populations reflecting the *DNMT3A*-mutant phenotype (Figure S 11).

As a second hypothesis, different temporal occurrence of the mutations could account for the differential global methylation between the outcome groups. The *DNMT3A* mutation event might have happened earlier in the ER samples. Thus, leading to an accumulation of more perturbations in the methylome and transcriptome, resulting in a more advanced form of AML. Desai et al. showed that pre-leukemic mutations often occur a decade before diagnosis in patients.[235] The temporal effect has been modeled in a study that introduced Dnmt3a-R878H (homolog to human *DNMT3A*-R882H) and *Npm1* mutations in mice with varying latency between events. Strikingly, increased latency led to a more aggressive disease with decreased overall survival in mice. In line with that, simultaneous mutations of Dnmt3a-R878H and Npm1 did not cause a lethal hematologic malignancy within 45 weeks in mice.[236,237] The authors hypothesized that accumulation of Dnmt3a-induced alterations synergizes with Npm1 mutations and causes an advanced aggressiveness. Other studies also highlight the effect of the temporal order of *DNMT3A* in combination with a *JAK2* mutation on hematologic malignancies.[238,239] More recent publications proposed that mutation of *DNMT3A* induced global methylation changes and particularly in HSCs triggered a selective advantage.[207,240]

Further studies observed transcriptomic instability with increased erroneous RNA splicing in *DNMT3A*-mutated cell lines.[124] Accordingly, differential splicing was also observed between outcome groups in the SyTASC cohort (Figure 33). Specific events with direct clinical implications (inclusion level differences > 50%, if not > 80%) could not be identified. Since the splicing machinery was not mutated or otherwise transcriptionally affected (data not shown), coordinated differences are probably not to be expected. The observed differences seem to manifest themselves mainly through a skewed distribution of event types, e.g., enhanced numbers of intron retention events in ER samples (Figure 33). In line with this observation,

reduction of alternative splicing and accordingly changes in specific event types has been previously described for cancer in general. Kim et al. observed higher levels of intron retention and lower levels of exon skipping in cancer compared to normal cells.[241] Interestingly, analysis of alternative splicing events suggested a significant co-occurance in genes with changes in DNA methylation (Table S 3). The effect of DNA methylation on alternative splicing has been oberserved in mutiple studies.[116] For example, Shukla et al. proposed a potential mechanism by modulation of elongation rates during transcription.[117] Taken together, these findings support the hypothesis that the enhanced hypomethylation observed in ER samples compared to LTR samples may reflect a temporal difference in the mutational events. Accumulating instability by progressing hypomethylation as a consequence of mutated *DNMT3A* might be associated to this process. Since specific mechanistic hypotheses have not been proposed yet, further research is needed to understand the priming of pre-leukemic HSCs. To investigate these probably subclonal effects, particularly single cell technologies will be helpful.

3.2.5 A trained outcome prediction signature is highly predictive in external AML cohorts

The data set was further used to train an outcome prediction signature for patient stratification. The signature was trained on the population-sorted data set together with the bulk RNA-sequencing data. This approach was chosen on one hand to account for LSC-enriched populations, which are most likely the origin of relapse, as well as on the other hand ensure applicability in external cohorts, which almost exclusively consist of bulk samples.[7] Additionally, this merging makes the training cohort bigger, and a larger cohort is better suited for machine learning, cross-validation, and avoiding overfitting of the training data set.[242] Instead of cox regression using RFS data, logistic binomial regression was chosen using the two outcome groups. The rationale behind this approach was the extreme difference in RFS. While ER samples relapsed within six months, some LTR showed relapse only after ten years. Therefore a more simple model was chosen also considering that relapse after multiple years might not be reflected in the transcriptional data at diagnosis anymore but rather might have originated from a subclonal or therapy-induced effect.

The signature showed overall high predictive power in external AML cohorts. Of the six tested cohorts, only two didn't reach significance, but still showed a trend. For these latter cohorts, the lower predictive power may be explained by very specific differences: the "TARGET AML"

is a pediatric cohort (while the SyTASC data used for training were obtained from adult AML patients), and the "Metzeler" cohort was transcriptionally analyzed via microarray analysis (while the SyTASC data underwent RNA-seq).[191,193] Also, the predictive power of the signature was better if cohorts were filtered for normal karyotype AML, indicating that the effect of complex cytogenetic rearrangements leads to processes not entirely represented in the SyTASC cohort. This observation might be a point of critisism, as the signature was trained on a genetically homogeneous cohort and a prediction for AML in general, may not necessarily be derived – or only if one assumes similar remission mechanisms in LSCs. Interestingly, filtering for only *DNMT3A*- and *NPM1*-mutated samples did not improve statistics (data not shown). This was possibly due to a low number of samples and other interfering mutations.

However, comparison with the stemness-related LSC17 signature revealed comparable, if not better predictive power. Furthermore, a combination of the outcome prediction signature with the LSC17 signature or the ELN classification indicated an additive effect. This suggested that the biological processes reflected by the signatures or the ELN classification are different but independently predict patient outcomes. Accordingly, genes in the trained outcome prediction signature and the LSC17 signature did not overlap.

Analysis of the 30 genes in the outcome prediction signature showed little shared biological function; among these, e.g., enrichment for regionalization or pattern specification could be identified. Some genes also showed differential methylation in the promoter region in the outcome groups of the SyTASC cohort. Particularly, *SORT1* expression displayed high predictive power reflected by its signature coefficient. Modarres et al. proposed *SORT1* to be involved in chemo-resistance in AML.[243] A fast relapse indeed indicates a chemo-resistance of the LSC populations. Hence, the outcome prediction signature might also be referred to as a chemo-resistance signature. In summary, the trained signature showed high predictive power in external cohorts and an added value compared to established classifiers.

3.2.6 Concluding remarks on results obtained from the SyTASC data set

One of the most striking advantages of the SyTASC cohort is the homogeneous genetic background. The leukemic phenotype highly depends on the exact mutations, and even different mutations within the same gene might cause different outcomes.[3,62,209] The enrichment of LSCs allowed to study biological processes facilitating the early relapse and thus most likely causing the chemo-resistant phenotype. The enrichment of LSCs by the novel sorting strategy could be additionally confirmed by the multi-omics data analyses and presented clear advantages compared to previous enrichment strategies for LSCs. Still, the sorted populations are enrichments of different cells, potentially including fast-cycling LSCs producing the leukemic progeny as well as more quiescent, therapy-persistent LSCs. Biological differences between the outcome groups were much more pronounced in these LSC-enriched populations compared to the more differentiated populations. As graphically summarized in Figure 57, ER samples showed a more stem-like phenotype potentially linked to a more quiescent subpopulation. This was reflected by the different energy states leading to a higher dependency on oxidative phosphorylation in the more stem-like, resistant ER samples; while more differentiated LTR samples presented a more glycolytic phenotype. DNA methylation and *NPM1* mutant allele frequencies led to the hypothesis of temporal differences between the outcome groups, potentially causing increased accumulated transcriptional instability. Evidence for a different cell of origin could not be found, and most likely, a posteriori processes led to the dedifferentiated, more stem-like phenotype in the ER samples. Furthermore, a predictive signature was trained that showed strong discriminatory power when compared to established prediction signatures – and the new signature was additive to the established ones potentially because it was liked to chemo-resistance mechanisms, an aspect not covered by the other expression signatures. In summary, important mechanisms could be identified which determine the relapse fate of patients. Nevertheless, further research is needed to study and validate these processes in vitro and in vivo. Notably, the differential expression of MHC-II genes and activity of TGFβ/BMP signaling requires investigation to understand the biological mechanism and identify potential therapeutic targets.

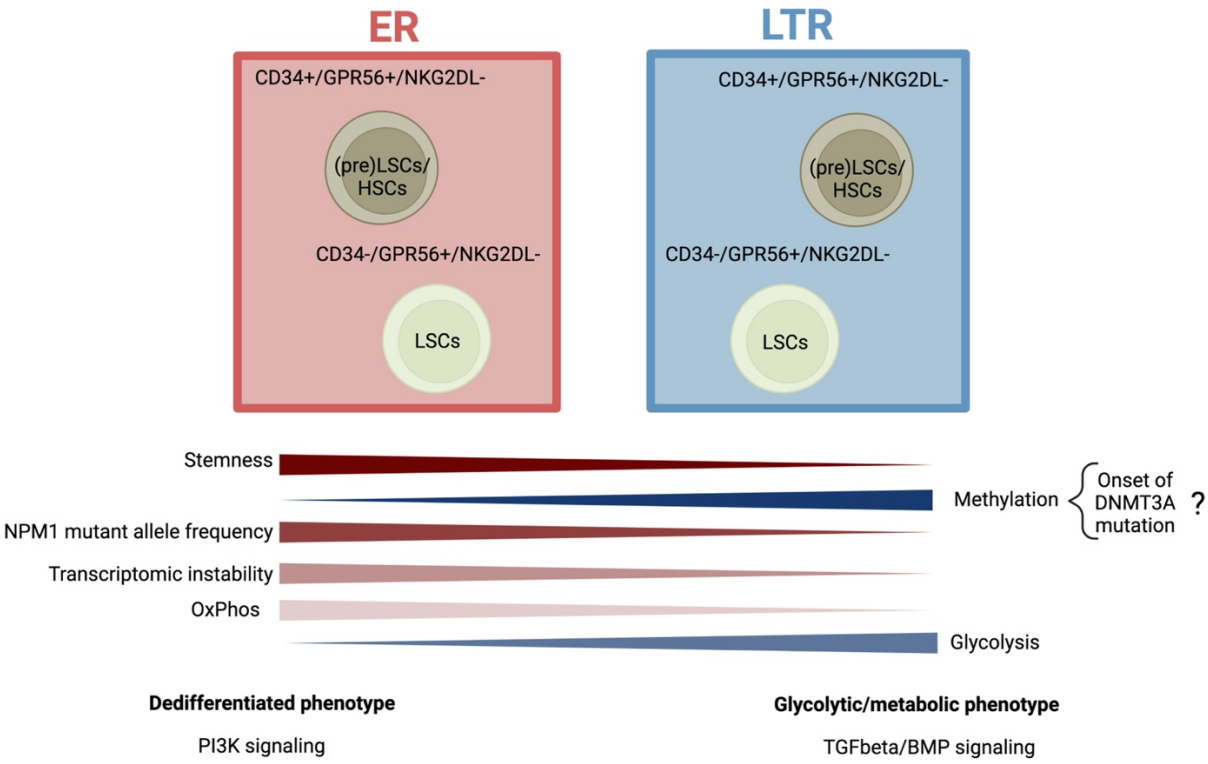

Figure 57: Overview of major differences between ER and LTR samples in engrafting LSC populations.

3.3 An interactive tool for metabolomics extraction protocol selection

Metabolic measurements are particularly informative since they represent the biological endpoint of transcriptional and regulatory processes. While the metabolic analysis of body fluids is well established, protocols for intracellular analysis are less standardized.[150–153] Especially for metabolomics, the pre-analytical phase, including sample collection and handling, is error-prone and has tremendous effects on the results. Compared to RNA and DNA, many metabolites are much less stable and have fast turnover in the cells.[155] These specifics must also be considered during the extraction of intracellular metabolites. The choice of an adequate extraction protocol for a given platform influences the range, robustness, and validity of the measurements.[244,245] The comparative study (see section 2.4) for ten different extraction methods on four human sample types (liver and bone marrow, as well as the cell lines HEK (adherent) and HL60 (non-adherent)) aimed to identify optimal protocols for metabolomics analyses. I could show that the QC metrics LOD, CV, and SOC reveal complementary information on the efficiency and repeatability of the measurements. The number of metabolites above the LOD gave an impression of the extraction efficiency, whereas the CV reflected the variability and repeatability for each metabolite for different sample types and protocols. The SOC was used as a metric for the global variability between replicates and revealed that some protocols are more prone to accumulation of technical variability during sample processing as it reflects the variation of multi-step experimental extraction protocols.

3.3.1 The optimal extraction protocol depends on sample type and metabolites of interest

Overall analysis showed that for liver samples the extraction efficiency was significantly higher than for the other tissue types. For liver samples, 30 mg of tissue was used as input material, whereas $3x10^6$ cells were used for HEK, HL60, and bone marrow samples. Comparing published values for liver cellularity, $(65 – 185) \cdot 10^6$ cells/g or $(139 \pm 25) \cdot 10^6$ cells/g have been described.[246,247] Thus, the used 30 mg of liver tissue corresponded to $(1.95 – 5.55) \cdot 10^6$ or $(4.17 \pm 0.75) \cdot 10^6$ cells, respectively. Therefore, the number of liver cells was in a similar range as for the other samples. Still, larger sample inputs probably increase extraction efficiency and repeatability.

Based on the QC metrics, the protocols *100 IPA* and *75 EtOH/MTBE B* showed the best results across the different sample types. This finding is in line with published comparisons in lipidomics and broad metabolic proofing; Calderón et al. showed that protocols containing isopropanol had good coverage, low technical variance, and absolute concentrations comparable to MTBE-based extraction protocols.[248] The technical advantage of isopropanol-based monophasic extractions is the lower volume of required solvents and laboratory scalability due to the rapid protocol which is potentially beneficial for stability and technical variance. With respect to coverage, MTBE- and chloroform-based extraction protocols were comparable. Notably, due to its toxicity, chloroform is a potential safety hazard, while MTBE is a non-toxic alternative that also showed less technical variance.[248,249]

The identification of the optimal method highly depended on the sample type and metabolites of interest. For liver tissue *75 EtOH/MTBE B*, *100 IPA,* and *100/30 IPA* resulted in the highest numbers of metabolites above LOD, respectively, allowing quantification of more than 400 metabolites. Hence, almost the full spectrum of the Biocrates MxP® Quant 500 kit could be covered. For bone marrow, *100 IPA* clearly performed best while *75 EtOH/MTBE B* and *EtOH/PP* also showed good results. However, for this tissue type, hardly more than 150 metabolites were effectively measured, thus only about one quarter of the potentially covered metabolites were found in detectable concentrations. For the HEK and HL-60 cell lines, *75 EtOH/MTBE B* showed by far the best yield of metabolites above the LOD covering slightly over 200 metabolites. Overall, extraction with protocol *MeOH/ACN/H2O* yielded the lowest number of detectable metabolites across all sample types, and generally, protocols containing methanol performed comparably weak. One reason for this might be that methanol is not nonpolar enough to extract the very nonpolar lipid species covered in this kit. In particular, clustering of PCA results showed that the solvents influence the profile of extracted metabolites. For the extraction of very hydrophobic lipid species, nonpolar solvents are required for effective quantification, e.g., isopropanol or MTBE which are used in classic lipidomics approaches.[250] In contrast, for the extraction of polar compounds e.g., certain amino acids, the influence of the protocol seems rather minor. The PCA also highlights the overall influence of the extraction protocols on the metabolite concentration patterns extracted from the same samples. Even though the kit allowed absolute quantification, comparability appeared to be limited to the same experimental settings. This further demonstrated the importance of thoughtful selection of the extraction protocol for larger studies.

3.3.2 An interactive tool for customized analysis

The comparative study showed that the optimal extraction protocol highly depends on the sample types and the chemical properties of measured metabolites. Therefore, the interactive R Shiny app "MetaboExtract" was implemented to explore and subset the data of this very data set in a tailored manner. The app allows to (de)select tissues, extraction protocols, and metabolite classes to focus on the data of interest and filter based on the LOD as well as the maximal CV. By applying this resource to identify a suitable protocol for the SyTASC data, I demonstrated the potential of this resource. Based on all metabolites, *75 EtOH/MTBE B* was identified as the protocol potentially yielding the best results. However, after filtering for lipids and adjusting the maximal variability, *MeOH/ACN/H2O* showed the best trade-off between range and repeatability, which was then also chosen for preprocessing for the metabolomics analysis of the SyTASC data. The app also includes a data set generated by Gegner et al. to identify optimal extraction methods for model organisms.[249] Additionally, the underlying R package "MetAlyzer", implemented by Nils Mechtel and me has been made available for data processing and conversion.

Besides the QC metrics covered in the app, it may be essential to consider the complexity of protocols and availability of chemical components for an optimal choice. The presented resources can be a valuable tool to support this crucial decision of an adequate extraction protocol for future metabolomics studies. The results can potentially be transferred to measurement technologies other than the Biocrates MxP® Quant 500 kit.

The results of this comparative study are presented in more detail in a joint publication published in *Frontiers Molecular Biosciences* (Andresen et al.[251]). In conclusion, the study provides a comprehensive comparison of different extraction methods for intracellular metabolic measurements and software for data processing as well as an online resource for interactive data exploration. Hence, it can improve the results of future studies while saving resources such as material and time.

4 Methods & Materials

4.1 Data sets

4.1.1 SyTASC cohort

The SyTASC consortium retrospectively selected the AML patient samples from the biorepositories AMLSG (ethics vote #148/10, ethics committee of the University of Ulm) and SAL (ethics vote EK98032010, ethical board of the Technical University Dresden). All patients carrying an *NPM1* (p.W288fs*12) and a *DNMT3A* (p.R882) mutation, received "7+3" chemotherapy and went into complete remission. Figure 58 shows an overview of the genetic profile of patients.

Healthy bone marrow samples were obtained from pseudonymized left-over samples with written informed consent in accordance with vote #329/10 (ethics committee of the Goethe University Medical Center Frankfurt). Processing of healthy samples was done as previously described.[252]

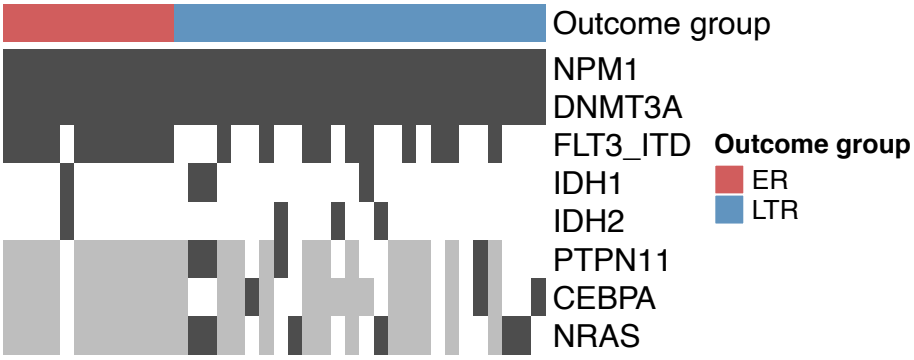

Figure 58: Heatmap showing the most frequently mutated genes in bulk AML samples stratified by ER and LTR. Genes were included if mutated in at least three samples. Dark grey: mutation, white: wt, and light grey: NA.

Characteristics and statistics of the SyTASC cohort and healthy bone marrow donors are shown in Table 5. Of note, high-throughput methods were performed on a subset of the 38 AML SyTASC patient samples. 37 samples were submitted to bulk RNA sequencing, samples from 31 patients were used for populations-based RNA-sequencing, and 26 samples for populations-based methylation analysis. Details on xenotransplantation results can also be found in Table S 1.

Table 5: Statistics of SyTASC cohort and healthy donor samples.

Parameter	SyTASC	Healthy
Number of samples	38	4
Sex [#]	m: 20 (52.6%); f: 18 (47.4%)	m: 2 (50.0%); f: 2 (50.0%)
Median age [years]	49 (22-66)	48.5 (41-54)
Median RFS [month]	40.2 (1.1-101.9)	-
Outcome group [#]	ER: 12 (31.6%); LTR: 26 (68.4%)	-
Median BM blasts [%]	76.5 (4-95.5), 1 NA	-
WBC [mio.]	62.75 (4.9-261.5)	-

4.1.2 External Data sets

External data sets used in this work are summarized in Table 6.

Table 6: Names and sources of external data sets. Gene Expression Omnibus (GEO).

Name	Type	Download/Availability	Publication
Oellerich	RNA-seq	EGA: EGAS00001005950	Jayavelu et al. (2022)[192]
TCGA	RNA-seq	http://firebrowse.org	Ley et al. (2013)[253]
GDC TCGA	RNA-seq	https://xenabrowser.net/[254]	Ley et al. (2013)[253]
GDC TARGET	RNA-seq	https://xenabrowser.net/[254]	Bolouri et al. (2018)[191]
Beat	RNA-seq	https://github.com/radivot/AMLbeatR	Tyner et al. (2018)[90]
Metzeler	Array	GEO accession: GSE12417	Metzeler et al. (2008)[193]
Corces	GSE12417	GEO accession: GSE74912	Corces et al. (2016)[20]
van Galen	scRNA-seq	GEO accession: GSE116256	van Galen et al. (2019)[40]
Jung	450k methylation	GEO accession: GSE63409	Jung et al. (2015)[196]

4.2 Experimental methods - SyTASC

All wet-lab experiments in context with the SyTASC samples were performed by Dr. Elisa Donato and Dr. Nadia Correia. Most experimental methods in this section have also been described in a manuscript submitted to *Blood Advances* (Donato*, Correia*, Andresen* et al., 2022, currently in revision cf. chapter “Own publications”).

4.2.1 Flow cytometry

For primary AML or healthy bone marrow samples, staining was performed in phosphate-buffered saline (PBS) + 2% fetal calf serum (FCS) using fluorescence conjugated antibodies targeting human CD3 (BioLegend), human CD19, human CD235a (eBiosciences), human CD20 (BD Biosciences), human CD34 (eBiosciences), human CD38 (Life Technologies), human GPR56 (BioLegend). Recombinant biotinylated NKG2D-Fc chimera was added to the mix to stain all NKG2D ligands.[98] DAPI (Sigma-Aldrich) or 7-AAD (BD Bioscience) were used for dead cell exclusion. Cells were FACS-sorted or analyzed using the gating strategy. Briefly, live cells were selected after the morphological gate and duplet exclusion. Therefore, SSC^{dim} $lineage^{low}$ cells were further selected with lineage positivity defined as $CD3^{+}CD20^{+}CD19^{+}CD235a^{+}$. Finally, five populations based on CD34, GPR56, and NKG2DL expression were defined: $CD34^{-}GPR56^{+}NKG2DL^{-}$, $CD34^{+}GPR56^{+}NKG2DL^{-}$, $CD34^{-}GPR56^{-}NKG2DL^{-}$, $CD34^{+}GPR56^{+}NKG2DL^{+}$, $CD34^{-}GPR56^{-}NKG2DL^{-}$ (cf. Figure 6). For the “NK-depleted” data set, additionally CD45 (BioLegend) was used to sort $Lineage^{Low}SSC^{dim}CD45^{dim}CD34^{-}GPR56^{+}NKG2DL^{-}$ and $Lineage^{Low}SSC^{dim}CD45^{dim}CD34^{+}GPR56^{+}NKG2DL^{-}$ populations. To check that NK cells were excluded from sorted populations, staining for NK cell markers CD94 (BD Bioscience) and CD56 (BioLegend) was performed.

PDX samples were stained in PBS + 2% FCS using anti-human CD45 (BD), anti-human CD33 (Life Technology), anti-human CD19 (eBioscience) and anti-murine CD45 (BioLegend) antibodies to evaluate the engraftment type (AML or multi-lineage). After morphological gate, exclusion of duplets and dead cells, PDX samples, were gated on $mCD45^{-}hCD45^{+}hCD33^{+}$ or $mCD45^{-}hCD45^{+}hCD19^{+}$.

Samples were FACS-sorted using BD FACS Aria Fusion (BD Biosciences) or analyzed using BD LSR Fortessa or BD LSR II (BD Biosciences).

4.2.2 Xenotransplantation assays

For xenotransplantation of bulk samples, primary patient samples were thawed, washed in Iscove's Modified Dulbecco's Medium (IMDM) (Gibco) + 10% FCS and treated with DNAse I (Roche) for 10 minutes at 37 °C. Cells were washed and $CD3^+$ cells depletion was performed according to the manufacturer's instructions using human CD3-conjugated microbeads from Miltenyi (130-050-101). $CD3^-$ cells were washed, counted, and transplanted intrafemorally in 8-12 weeks old female NSG mice one day after sub-lethal irradiation (175 cGy).

For xenotransplantation of populations sorted samples, thawed primary samples were washed in IMDM + 10% FCS and treated with DNAse I (Roche) for 10 minutes at 37 °C. Probes were then washed, stained, and sorted as described in the subsection "Flow cytometry". Sorted populations were resuspended in PBS and between $2.2x10^3$ and $1x10^5$ cells per mouse were transplanted intrafemorally in 8-12 weeks old female NSG mice one day after sub-lethal irradiation (175 cGy).

Human engraftment was monitored for 8 to 50 weeks by peripheral blood sampling or bone marrow aspiration (every 2 weeks or 6 weeks, respectively) or at signs of distress. Engraftment was defined as $\geq$ 0.1% human cells in murine bone marrow measured by flow cytometry (BD LSR Fortessa or BD LSR II, BD Biosciences) using anti-human CD45 (BD), anti-human CD33 (Life Technology), anti-human CD19 (eBioscience) and anti-murine CD45 antibodies (BioLegend).

NOD/SCID/IL2Rγ^{null} (NSG) mice (Jackson Laboratory) were hosted under pathogen-free conditions according to the German federal and state regulations (approved by the Regierungspräsidium Karlsruhe, Tierversuchsantrag number G43/18 and Z110/02).

4.2.3 RNA extraction

Thawed primary samples were washed in IMDM + 10% FCS and treated with DNAse I (Roche) for 10 minutes at 37 °C. Probes were then washed, stained, and sorted as described in the subsection "Flow cytometry". $1x10^3$ to $50x10^3$ cells from each population were directly sorted into RNA lysis buffer, and RNA was isolated using the Arcturus PicoPure RNA Isolation Kit (Life Technologies) according to the manufacturer's instructions. RNA integrity was checked with Bioanalyzer using Agilent RNA 6000 Pico Reagents (Agilent).

4.2.4 RNA-seq library preparation

Extracted RNA was converted into cDNA using the SMART Seq v4 ultra-low RNA kit (Takara). Libraries were produced using NEBNext® ChIP-Seq Library Prep Master Mix Set for Illumina® (New England Biolabs). Libraries were sequenced in paired-end mode with a read length of 125 base pairs (bp) on the HiSeq2000 V4 (Illumina) system. The "NK-depleted" samples were pre-pared analogously but sequenced on the NextSeq 550 (Illumina) system.

Of note, RNA-seq was performed in two batches. Hence, a prominent batch effect was observed, which was accounted for as described in the respective computational downstream analyses. If not otherwise indicated, of those samples sequenced in both batches, only samples from the "first" batch were included (Figure S 1).

4.2.5 Methylation profiling

DNA extraction was performed according to the manufacturer's instructions using the QIAamp DNA Micro Kit (QIAGEN). DNA was quantified using Qubit dsDNA HS Assay Kit (Life Technologies). DNA methylation profiling was performed using the Infinium® MethylationEPIC Kit (Illumina) at the Genomics and Proteomics Core Facility of the German Cancer Research Center (DKFZ).

4.2.6 Metabolomics

For metabolomic analysis, bones from primary PDX mice (3 biological replicates for each patient sample) were collected and crashed in RPMI. Bone marrow cells were collected, and red blood cells were lysed using an ammonium-chloride-potassium (ACK) lysis puffer. Cells were then washed and stained in RPMI with fluorescence conjugated antibodies: anti-human CD45 (BD), anti-human CD34 (eBiosciences), anti-human CD38 (BD) anti-human GPR56 (BioLegend), anti-human CD33 (Life Technologies), anti-human CD19 (eBioscience), anti-human CD64 (BioLegend), and anti-murine CD45 antibodies (BioLegend). 7-AAD (BD Bioscience) staining was used for dead cell discrimination. After morphological gate and exclusion of duplets and dead cells, $mCD45^{-}hCD45^{+}hCD34^{+}hGPR56^{+}$ and $mCD45^{-}hCD45^{+}hCD34^{-}hGPR56^{+}$ were FACS-sorted in RPMI + FCS. Cells were then spun down, washed with ammonium carbonate (75 mM, pH 7.4), spun down, and the dry pellet was snap-frozen in liquid nitrogen.

Metabolite extraction was performed by adding 50 µl of cold extraction solvent (40:40:20 acetonitrile:methanol:water). After incubation for 20 minutes at -20 °C, cells were vortexed and spun down. The supernatant was collected and used for metabolite quantification.

Untargeted metabolomics was performed by Prof. Dr. Nicola Zamboni by flow injection analysis (FIA) on an Agilent 6550 iFunnel Q-TOF LC/MS instrument in duplicates as previously described at the Institute of Molecular Systems Biology of the ETH Zurich.[255] An unpublished in-house software was used to perform peak integration of mass spectra and processing, including visualization of covered metabolites.

4.3 Experimental methods – Extraction comparison

Experimental methods in this section have also been published as part of a paper in *Frontiers in Molecular Sciences* (Andresen et al.[251]). Sample collection and processing were performed by Dr. med. Tobias Boch and Andreas Narr. Mass spectrometry analysis was run by Dr. Gernot Poschet and Dr. Hagen M. Gegner.

4.3.1 Sample preparation

All primary human samples (liver and bone marrow) in this study were obtained and used following institutional review board approval by the Medical Ethics Committee II of the Medical Faculty Mannheim, University of Heidelberg, Germany, in accordance with the declaration of Helsinki after informed written consent.

Liver tissue from a 75-year-old male patient with hepatocellular carcinoma was obtained during surgical liver resection. After surgical resection, part of the healthy liver tissue was immediately washed with ice-cold 0.9% NaCl solution and snap-frozen in liquid nitrogen. The frozen tissue was pulverized to a fine powder using a ball mill (2x 30 sec; 30 Hz; MM 400; Retsch) and pre-cooled stainless-steel beakers. Aliquots of 30 mg were stored at -80 °C until extraction.

The bone marrow was obtained by aspiration from a healthy 39-year-old male volunteer and was performed according to standard clinical protocols. Mononuclear cells (MNCs) were isolated from fresh bone marrow by Ficoll density gradient centrifugation. To do so, bone marrow was diluted 1:2 with PBS and loaded on the Ficoll without disturbing the layer.

Samples were centrifuged at 400 x g at room temperature for 30 minutes. MNCs were extracted and washed twice with PBS. Aliquots of $3x10^6$ cells were collected and snap-frozen using liquid nitrogen.

The two cell lines, adherent human embryonic kidney (HEK) and non-adherent human leukemia 60 (HL60), were kept under cell culture conditions in Dulbecco's Modified Eagle's Medium (DMEM) GlutaMAX (Gibco) with 10% FCS and 1% Penicillin/Streptomycin. Cells were washed twice with ice-cold PBS and aliquots of $3x10^6$ cells were collected and snap-frozen using liquid nitrogen.

4.3.2 Metabolite extraction

The ten extraction protocols for intracellular metabolomics measurements are outlined in (Figure 59). Briefly, metabolites from the frozen cell ($3x10^6$) or liver tissue (30 mg) samples were extracted using the respective solvents and subsequent steps of the different protocols. For the ultra-sonification ice bath Transsonic T460 (Elma) was used. After a final centrifugation step, the extraction solvents of the protocols *100/30 IPA*, *100 IPA*, *MeOH/ACN/H2O*, *MeOH/ACN*, and *EtOH/PP* was transferred into a new 1.5ml tube (Eppendorf) and snap-frozen until kit preparation. For all other protocols, the supernatant (in the biphasic extractions with MTBE or chloroform both phases) was dried using an Eppendorf Concentrator Plus set to no heat. Samples were stored at -80°C and reconstituted in 120 or 60 µl isopropanol (60 or 30 µl of 100% isopropanol, followed by 60 or 30 µl of 30% isopropanol in water) directly before measurement. The protocol *75 EtOH/MTBE* was applied with two resolving volumes "A" 120 µl and "B" 60 µl (Figure 59). All chemicals and solvents used were of UHPLC-MS grade quality (Sigma-Aldrich).

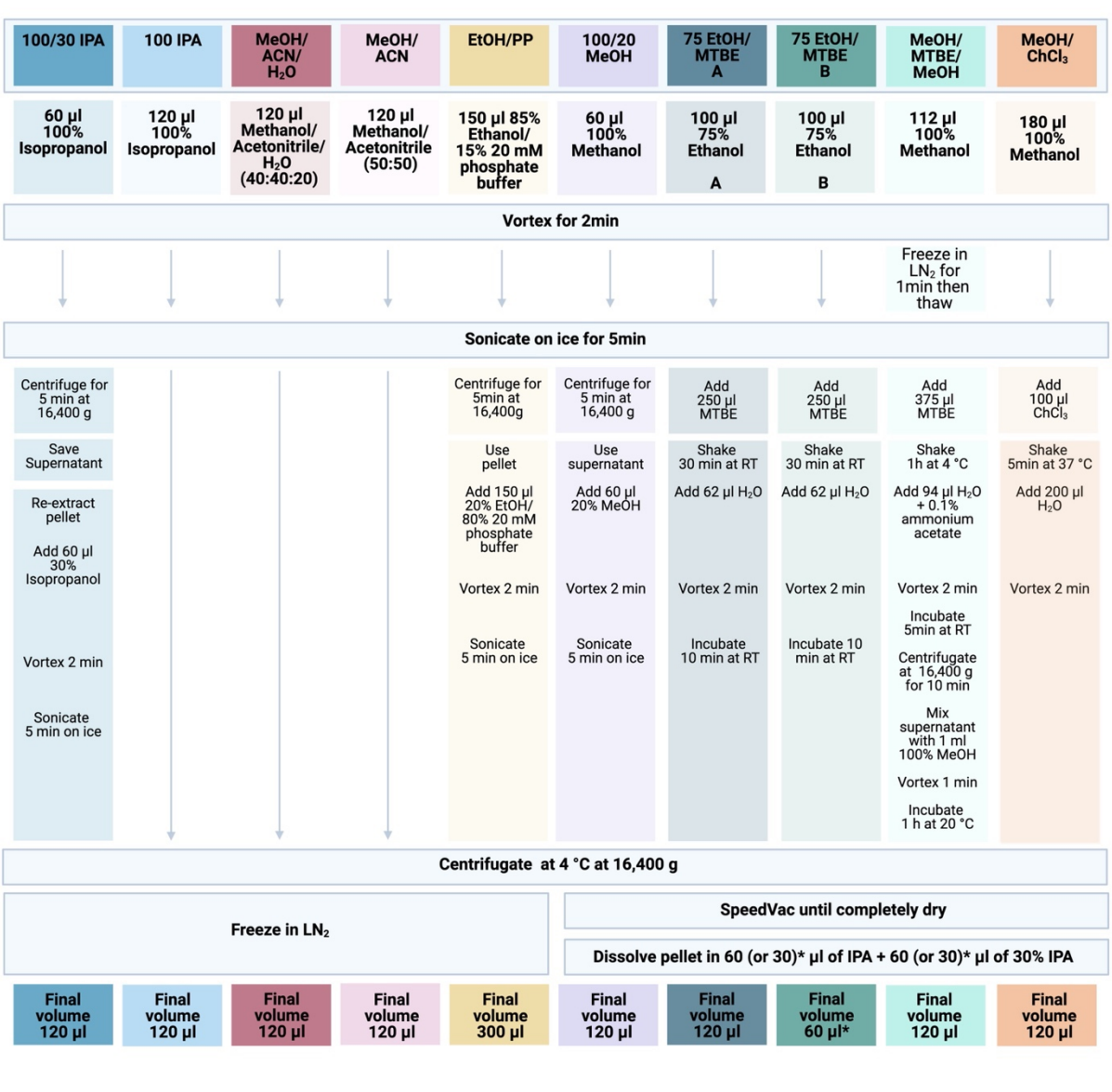

Figure 59: Extraction protocols for comparative metabolomics study. IPA: isopropanol; MTBE: methyl tert-butyl ether; ACN: acetonitrile; EtOH: ethanol; MeOH: methanol; PP: polypropylene; $ChCl_3$: chloroform; LN_2: liquid nitrogen; RT: room temperature. (Adapted from Andresen et al.[251])

4.3.3 Sample analysis

In total, 630 metabolites covering 14 small molecule and 9 different lipid classes were analyzed using the MxP® Quant 500 kit (Biocrates) following the manufacturer's protocol on a SCIEX QTRAP 6500+ mass spectrometry system. As previously described in my publication (Andresen et al.[251]), measurements were performed by Dr. Gernot Poschet and Dr. Hagen M. Gegner at the Metabolomics Core Technology Platform core facility of the Heidelberg University. Data were recorded using the Analyst software suite (Sciex) and transferred to the MetIDQ™ software (version Oxygen-DB110-3005), which was used for further technical validation, quantification, and data export.

4.4 Computational Methods

4.4.1 RNA-seq alignment

RNA-seq alignment was performed by the DKFZ in-house One Touch Pipeline which implements the RNA-seq pipeline of the DKFZ Omics IT and Data Management Core Facility (https://github.com/DKFZ-ODCF/RNAseqWorkflow). Briefly, base calling was performed using Bcl2fastq2 2.20. Reads were trimmed for adapter sequences and aligned to the 1000 Genomes Phase 2 assembly of the Genome Reference Consortium human genome (build 37, version hs37d5) with STAR (v2.5.3a)[256]. For alignment, the following parameters were used: alignIntronMax: 500.000, alignMatesGapMax: 500.000, outSAMunmapped: Within, outFilterMultimapNmax: 1, chimScoreMin: 1, outFilterMismatchNmax: 3, sjdbOverhang: 50, outFilterMismatchNoverLmax: 0.3, chimSegmentMin: 15, chimScoreJunctionNonGTAG: 0, chimJunctionOverhangMin: 15. For building the index, GENCODE gene annotation (GENCODE Release 19) was used.[257] BAM files were generated using SAMtools (v1.6).[258] Duplicates were marked with Sambamba (v0.6.5)[259] and raw counts were generated using featureCounts (Subread version 1.5.3).[260]

4.4.2 RNAseq analysis

For calculation of transcripts per million (TPM), mitochondrial RNA, transfer RNA, ribosomal ribonucleic acid RNA, as well as all transcripts from the Y- and X-chromosome were removed. Subsequently, normalization was performed analogously to standard TPM calculation.

Clustering analysis

For downstream analysis, counts were processes using a variance-stabilizing transformation (vst) as implemented in DESeq2[261] and corrected for batch effects using the function "rescaleBatches" implemented in the R package batchelor[262] (cf. Figure S 1). PCA was performed on the top 500 variable genes using the function `prcomp()` with `center = T` and `scale = T`. Consensus partitioning was formed using the R package cola. Standard deviation (SD) was identified as the best top-value method and hierarchical clustering together with `cutree()` (hclust) was identified as the best partitioning method. A partition number of 2 was identified as optimal after iteration over `k_max = 6`.

Differential expression and GSEA

Analysis of differential expression between samples was performed using DESeq2.[261] The batch information was always included in the design term. The function `lfcshrink()` was used to define differentially expressed genes (`abs(log2FC) ≥ 1, p.adj ≤ 0.05`)[263]. Log2 fold changes (non-shrinked) were used for GSEA analysis with clusterProfiler[264] and the Molecular Signatures Database v7.4.1[265] as reference gene sets and `pvalueCutoff` set to 0.25. GSEA plots were generated using the function `gseaplot2()` implemented in the R package enrichplot.

For plotting of expression values, count data was divided by the size factor calculated by the function `estimateSizeFactorsForMatrix()` implemented in the R package DESeq2, log2-transformed and batch-corrected as described in the subsection "Clustering analysis". These expression values are referred to as "log2(normCounts)".

LSC signatures

LSC17 signature score was calculated on log2-transformed TPMs. Correlation with the 104-gene LSC signature was calculated on log2-transformed TPMs using the "spearman" method.[51]

Cell cycle analysis

To estimate the cell cycle phase, functions implemented in the R package Seurat were used.[266] Count data were normalized by the function `NormalizeData()`. Variable features were identified using the function `FindVariableFeatures()` with `selection.method = "vst"` and scaled using the function `ScaleData()`. Cell cycle phase, S and G2M scores were estimated by the function `CellCycleScoring()`.

Analysis of intronic and exonic counts

Intronic and exonic read counts were analyzed by implementing the exon-intron split analysis (EISA) described by Gaidatzis et al. [180] based on featureCounts (Subread version 1.5.3). Read counting was set to "unstranded" and reads that overlapped regions of multiple genes were discarded. Counts were calculated for exon and gene bodies (parameter -t set to "exon" or "gene"). As described by Gaidatzis et al., exonic coordinates were extended by ten base pairs on both sides in the input gtf file (GENCODE Release 19, cf. "RNA-seq alignment"). This procedure ensures that exonic reads close to the exon junction were not counted as intronic.

To calculate intronic counts, exonic counts were subtracted from gene body counts. If intronic counts were negative, values were adjusted to zero.

Embedding into healthy hematopoiesis

For the embedding of samples into healthy hematopoiesis, the population-sorted RNA-seq data were integrated with data from the healthy samples from the van Galen data set using the R package Seurat.[40,266] Data sets were normalized, variable features were identified and scaled separately (`NormalizeData()`, `FindVariableFeatures()` and `ScaleData()`). For integration, `SelectIntegrationFeatures()`, `FindIntegrationAnchors()`, `IntegrateData()` and `ScaleData()` was used with `dims = 1:20`, `reduction = "rpca"` and `k.weight = 43`. The integrated data set was subjected to the functions `RunPCA()` and `RunUMAP()` with `dims = 1:20`. Positions of labels indicate the mean of the respective cell types as indicated in the van Galen data. To calculate the distance between HSC samples and SyTASC samples, the Euclidian distance between the median positions in the UMAP was determined.

Estimation of HSC proportions

As a reference for healthy hematopoietic cells, the Corces data set was used. Corces and SyTASC sorted populations were integrated by the R package Seurat as described above setting `dims = 1:10`, `npcs = 10`, `k.score = 20` and `k.weight = 10`. Non-negative least squares implemented in the R package nnls was used to deconvolve the sorted SyTASC populations with median expression of normalized and integrated Corces data as cell type-specific reference expression signatures.[267]

Transcription factor and pathway activity

Virtual Inference of Protein-activity by Enriched Regulon (VIPER) was used to estimate transcription factor activity by the wrapper function `run_viper()` implemented in the R package dorothea based on the viper algorithm.[268–270] As a reference, regulons published in the R annotation package hs_dorothea by Holland et al. were used.[269] The parameter `minsize` was set to 5 and the `stat` result estimated by differential expression analysis via DESeq2 was used as input for the respective comparison. Volcano plots were created with the function `volcano_nice()`.

Pathway activities were estimated by the Pathway RespOnsive GENes for activity inference (PROGENy) method implemented in the R package progeny, again using the `stat` result from DESeq2 described above as input.[271] The parameter `perm` was set to 10,000. The expression of pathway-related genes was plotted using the function `scat_plots()`.

Alternative Splicing

Alternative splicing events between outcome groups were analyzed with the software rMATS v 4.1.1 installed in a conda environment with anaconda3 (version 2019.07). The script `rmats.py` was run with the following settings in python version 3.7: Paired-end mode was chosen with a read length of 125 bp and `novelSS` was used to allow the identification of novel splice sites. Analysis was performed using BAM files as input. Significant events were filtered using a false discovery rate (FDR) ≤ 0.05. Alternative splicing analysis was only performed for the larger "second" sequencing batch (cf. Figure S 1).

Allele frequencies

Mutant allele frequencies from RNA-seq data were determined using the `count()` function implemented in igvtools (version 2.4.14).[272] For DNMT3A (p.882) mutations position chr2:25457242-25457243 and for the NPM1 (p.W288fs*12) mutation position chr5:170837547-170837548 were inspected in BAM files.

4.4.3 LASSO regression

For the identification of genes discriminating between the two outcome groups, least absolute shrinkage and selection operator (LASSO) regression implemented in the R package glmnet was used in logistic regression "binomial" mode.[273] All SyTASC samples from bulk and population-sorted sequencing were included. TPMs were used as input. Different batches were integrated by fitting a linear model for all genes separately for the overlapping samples and used to adjust the values of the "second" batch accordingly. Negative values were set to zero. Figure S 1 shows the concept and validation of the batch correction.

Additionally, the input data was filtered for "protein-coding" genes according to the annotation gtf file (cf. subsection "RNA-seq alignment"). Only genes with TPM ≥ 10 in all samples were included and log2-transformed. To avoid overfitting, the function `cv.glmnet()` was used to identify the optimal λ with and 10-fold cross-validation and the intercept was set to zero. 100 cross-validations were performed and the optimal λ was chosen as a trade-off between the minimal mean cross-validated error, a low number of coefficients

and the calculated F1 score for the training cohort (Figure S 15). The signature was then fit with the function `glmnet()` with intercept set to zero.

For external validation, the signature score was calculated for each sample on log2-transformed values provided by the authors of the data sets (Oellerich: TPM, TCGA: Reads Per Kilobase Million (RPKM), GDC TCGA: Fragments Per Kilobase Million (FPKM), GDC TARGET: FPKM, Beat: RPKM, Metzeler: microarray output, SyTASC: TPM) (cf. Table 6). The median signature score for each cohort was used to stratify the cohort into "low" and "high". These strata were used to calculate Kaplan-Meier estimators and log-rank tests were performed to estimate the significance as implemented in the R package survival.[274,275] Plots were generated using the function `ggsurvplot()` implemented in the R package survminer.[276]

4.4.4 Methylation analysis

Processing

Raw methylation data was imported and processed using the R package minfi[277] and annotated by the R annotation package IlluminaHumanMethylationEPICanno.ilm10b2.hg19. The function `preprocessIllumina()` was used for processing and normalization. In addition, probes covering single-nucleotide polymorphism (SNP) loci were removed.

Clustering analysis

PCA was performed on the top 1000 variable genes using the function `prcomp()` with `center = T` and `scale = T`. Consensus partitioning was formed using the R package cola. Standard deviation (SD) was identified as the best top-value method, and k-means clustering (`kmeans`) was identified as the best partition method. A partition number of 2 was identified as optimal after iteration over `k_max = 6`.

Enrichment analysis of genomic regions

Enrichment analysis of the most variable genomic regions was performed by LOLA.[278] This algorithm identifies enrichment for transcription factors interacting with the queried regions. Regions for all measured probes were used as the `userUniverse` and the CODEX database was used as the reference.[279] For over-representation analysis, transcription factors were filtered for qvalues ≤ 0.05. Over-representation analysis was performed using the function

enrichGO() implemented in the R package clusterProfiler with ont = "ALL". Results were visualized by the function dotplot() implemented in the R package enrichplot.[280]

Differentially methylated positions and regions

For differential analysis, the R package minfi was used. Differentially methylated positions (DMPs) were calculated using the function dmpFinder() with type = "categorical". Differentially methylated regions (DMRs) were calculated using the function bumphunter() with cutoff = 0.15 and B = 1000. Significance was defined by an absolute beta value difference > 0.25 and family-wise error rate (FWER) < 0.05. DMPs were visualized using the R package Gviz and the annotation package TxDb.Hsapiens.UCSC.hg19.knownGene as genomic reference.

Cell-of-origin of cancer

Jung et al. published 216 DMRs specific for HSCPs.[196] The cell of origin was estimated by clustering of beta values of probes in these DMRs. Not all probes could be matched because of differences in the kits. SyTASC data were profiled with the Infinium® MethylationEPIC Kit (Illumina), while the data published by Jung et al. were profiled using the Infinium® HumanMethylation450 BeadChip (Illumina). Withing the 216 regions, the euclidean distance for beta values of 455 probes was calculated and subsequent hierarchical clustering was performed before visualization with the R package dendextend.[281]

4.4.5 Data integration

For the integration of transcriptomic, methylation, and mutation data from population-sorted SyTASC AML samples, the R package MOFA was used.[282] Transcriptomic data was normalized and batch-corrected as described in the "Clustering analysis" subsection. In addition, genes with an estimated variance < 0.1 were removed from the data set. Illumina normalized methylation data were used and probes with an estimated variance < 0.1 were removed. Moreover, probes with only "non-available" information were removed. Mutation information was included as a binary matrix analogous to Figure 58. Of note, genes were included if mutated in at least three and non-mutated in at least three samples. Model options were adjusted to numFactors = 30. Train options were set to DropFactorThreshold = 0.01, tolerance = 0.01, and maxiter = 1000. MOFA was run with ten different seeds and the optimal run was chosen using the function

`selectModel()`. The feature overview was plotted with the function `plotDataOverview()` and the explained variances were plotted with the function `plotVarianceExplained()`. GSEA was performed as described in the subsection "Differential expression and GSEA" but with loadings of LFs as input.

4.4.6 Untargeted metabolomics

After pre-processing (see above), the data were manually curated aiming to remove empty injections and outlier samples. The curated data was mean normalized and further analyzed in R version 4.0.0. Additionally, batch effects between biological replicates were corrected using the function `ComBat()` implemented in the R package sva. Differentially abundant metabolites were determined by fitting a linear model on log2-transformed values. Metabolite abundance was used as the response and the outcome group as the independent variable. Significant was calculated by a Wald chi-squared test using the variance-covariance matrix and the coefficients of the fitted model. qvalues were calculated using the R package qvalue to account for multiple testing. Significance was considered for abs(log2FC) > 0.2 and qvalue < 0.05.

Enriched metabolic pathways were calculated based on Human Metabolome Database (HMDB) pathway version 3.0. Significance was estimated by a Fisher's exact test for significantly higher and lower abundance of metabolites between the outcome groups. All measured metabolites were grouped in the contingency table by presence in the respective pathway and significance.

4.4.7 MetAlyzer, MetaboExtract, and extraction comparison analysis

Together with Nils Mechtel, I developed the R package MetAlyzer and the R Shiny app in R version 4.0.0. MetAlyzer is available on the open repository CRAN (https://CRAN.R-project.org/package=MetAlyzer). The code is also documented along with the most recent software version on GitHub (https://github.com/andresenc/MetAlyzer). MetaboExtract is accessible online (http://www.metaboextract.shiny.dkfz.de). The underlying code is available on GitHub (https://github.com/andresenc/MetaboExtract).

The metabolomics analyses presented in this work can be analogously generated by MetaboExtract, except for the PCA. Metabolites were defined as "above LOD" when at least 2 of the 3 replicates met this criterion. Prior to the analysis in MetaboExtract, LODs were

calculated by the MetIDQ™ software. For PCA, metabolites below LOD were filtered out from raw data in all samples. Zero values were replaced by taking the minimum of all measured concentrations per metabolite and adding 20% to this value. Filtered data were log2-transformed and scaled using the Pareto method.[283] To identify associations between principal components and the experimental setting, Kruskal-Wallis rank sum tests were applied. Significant was defined by p-values < 0.05.

4.4.8 Visualization

If not otherwise indicated, figures were created using the R packages graphics, ggplot2, ComplexHeatmap, and igraph. Figures 1-6, 56,57, and 59 were created with BioRender.com. Figures were assembled and visually optimized in Adobe Illustrator version 26.0.2.

4.4.9 Statistics

Statistical testing was performed as described in the respective figure captions, including significance levels.

The function coxph() implemented in the R package survival was used to fit the Cox proportional hazard regression model. [274,275]

4.5 Software and Packages

All computational analyses were run under CentOS Linux 7 (Core) cluster environment managed by the Omics IT and Data Management Core facility. Details on R packages and annotation packages are summarized in Table 7 and Table 8. R version 3.6.0 and RStudio version 1.4 was mainly used except for operations that required import of the packages Seurat, dorothea and progeny, as well as the development of MetaboExtract and MetAlyzer. These were run using R version 4.0.0.

Table 7: R annotation packages and versions used in this work.

Package	v3.6.0
BSgenome.Hsapiens.UCSC.hg19	1.4.0
FlowSorted.Blood.450k	1.22.0
FDb.InfiniumMethylation.hg19	2.2.0
hgu133plus2.db	3.2.3
IlluminaHumanMethylation450kanno.ilmn12.hg19	0.6.0
IlluminaHumanMethylationEPICanno.ilm10b2.hg19	0.6.0
TxDb.Hsapiens.UCSC.hg19.knownGene	3.2.2

Table 8: R packages and versions in different R versions used in this work.

Package	v3.6.0	v4.0.0	Package	v3.6.0	v4.0.0
agricolae	1.3-5	-	magrittr	2.0.1	2.0.1
AnnotationDbi	1.46.1	-	Matrix	1.2-17	-
aod	1.3.1	-	matrixStats	0.58.0	0.61.0
apeglm	1.6.0	-	methylumi	2.30.0	-
batchelor	1.0.1	-	mgcv	-	1.8-40
Biobase	2.44.0	2.48.0	minfi	1.30.0	-
BiocGenerics	0.32.0	0.34.0	MLmetrics	1.1.1	-
BiocParallel	1.18.1	1.22.0	MOFA	1.0.0	-
biomaRt	2.40.5	2.44.4	msigdbr	7.4.1	-
Biostrings	2.52.0	-	MultiAssayExperiment	1.10.4	1.14.0
BSgenome	1.52.0	-	nlme	-	3.1-157
Bumphunter	1.26.0	-	nnls	1.4	1.4
CePa	0.7.0	-	org.Hs.eg.db	3.8.2	-
circlize	0.4.12	0.4.13	plotly	4.9.3	4.9.4.1
ClusterProfiler	3.12.0	-	progeny	1.6.0	1.10.0
cola	1.0.1	-	purr	0.3.4	0.3.4
ComplexHeatmap	2.0.0	2.4.3	qvalue	2.16.0	2.20.0
Cowplot	1.1.1	1.1.1	RColorBrewer	1.1-2	-
data.table	1.14.0	1.14.0	Rcpp	1.0.6	-
DelayedArray	1.10.0	0.14.1	reporttools	1.1.2	-
dendextend	1.15.1	-	reshape2	1.4.4	-
DESeq2	1.24.0	1.28.1	ROC	1.60.0	-
devtools	2.4.1	-	rtracklayer	1.44.4	-
dorothea	-	1.0.1	Rtsne	-	0.15
dplyr	1.0.6	1.0.7	S4Vectors	0.22.1	0.26.1
enrichplot	1.4.0	-	scales	1.1.1	1.1.1
fgsea	1.10.1	1.14.0	Seurat	-	4.1.0
genefilter	-	1.70.0	simpleCache	0.4.2	-
GenomeInfoDb	1.20.0	1.24.2	SingleCellExperiment	1.6.0	-
GenomicFeatures	1.36.4	-	stringi	1.6.2	-
GenomicRanges	1.36.1	1.40.0	stringr	1.4.0	1.4.0
ggplot2	3.3.3	3.3.5	SummarizedExperiment	1.14.1	1.18.2
ggrepel	0.9.1	0.9.1	survival	3.2-11	-
ggsignif	0.6.1	0.6.3	survminer	0.4.9	-
glmnet	4.1-1	-	sva	-	3.36.0
Gviz	1.28.3	-	tibble	3.1.2	3.1.4
igraph	1.2.6	-	tidyr	1.1.3	1.1.3
illuminaio	0.26.0	-	tidytable	0.6.2	0.6.5
IRanges	2.18.3	2.22.2	umap	-	0.2.7.0
limma	3.40.6	-	viper	1.18.1	1.22.0
locfit	1.5-9.4	-	xlsx	0.6.5	-
LOLA	1.14.0	-	xtable	1.8-4	-
lumi	2.36.0	-	XVector	0.24.0	-

5 Contributions

All wet lab experiments in context with the SyTASC samples were performed by Dr. Elisa Donato and Dr. Nadia Correia, who also developed the cell sorting strategy.

SyTASC samples were selected by the SyTASC consortium (Prof. Dr. Frank Buchholz, Prof. Dr. med. Lars Bullinger, Prof. Dr. med. Christian Thiede, Prof. Dr. Andreas Trumpp, Prof. Dr. Michael A. Rieger, Prof. Dr. med. Hubert Serve, Prof. Dr. med. Daniela Krause, and Prof. Dr. Ingo Roeder) funded by the “Deutsche Krebshilfe”.

Untargeted metabolomics analysis was performed by Prof. Dr. Nicola Zamboni, who also provided software, infrastructure, and scientific expertise for computational analyses.

Eva Holtkamp (student in Molecular Biotechnology at Heidelberg University whom I co-supervised during her bachelor’s thesis and a Master internship) implemented the EISA method and contributed to the bulk sample deconvolution approaches. Kamil Moskal (student in Molecular Biosciences at Heidelberg University whom I co-supervised during a Master internship) implemented the rMATS workflow for alternative splicing analysis and Ferdinand Popp (student in the Major in Cancer Biology at DKFZ and Heidelberg University whom I co-supervised during a Master internship) supported training of the signature by LASSO regression.

RNA-seq alignment was performed at the DKFZ Omics IT and Data Management Core Facility.

Sample collection and processing for the metabolomics extraction protocol evaluation were performed by Dr. med. Tobias Boch and Andreas Narr. Mass spectrometry analysis was performed at the Metabolomics Core Technology Platform (MCTP) core facility of the Heidelberg University by Dr. Gernot Poschet and Dr. Hagen Gegner, who also provided scientific expertise for metabolomics analysis.

The R package MetAlyzer and the R shiny app MetaboExtract were implemented together with Nils Mechtel.

Scientific guidance throughout the projects was provided by Dr. Dr. med. Daniel Hübschmann and Prof. Dr. Andreas Trumpp, who were essential for the project's shape and direction.

This thesis has been proof-read and corrected by Dr. Dr. med. Daniel Hübschmann.

6 Acknowledgement

This work would not have been possible without the support of many great people who helped and guided me. The past four years have been an exciting, yet challenging journey. It gave me multiple opportunities to develop as a scientist and also as a person, and I am very grateful to so many inspiring people I have met along this way.

First of all, I would like to thank my supervisor **Daniel Hübschmann**, who initially inspired me to pursue bioinformatics. Thank you for your supervision, your support, ideas, and advice. Thank you for all the freedom and opportunities throughout my PhD, for motivating me, always being optimistic, and for sharing your tremendous scientific knowledge.

I would also like to thank **Andreas Trumpp** for giving me the chance to work in his lab. Thank you for creating such a unique place at the DKFZ. I really enjoyed the positive and inspiring atmosphere at HI-STEM. Thank you for all your ideas, your scientific input and enthusiasm throughout my PhD, and all the freedom and opportunities including courses, conferences, and retreats.

I would also like to thank **Prof. Dr. Jan Lohmann** and **Dr. Christiane Opitz** for being part of my thesis examination committee.

Further, I would like to thank **Dr. Carl Herrmann** and **Dr. Wolfgang Huber** for the scientific discussions and helpful input during my thesis advisory committee meetings.

I like to thank **Elisa Donato** and **Nadia Correia** for their work on the SyTASC project and scientific discussions.

Thank you, **Gernot Poschet** and **Hagen Gegner** for your enthusiasm for the metabolomics project, ideas, scientific discussions, and your expertise. Besides, I would like to thank **Tobias Boch** and **Andreas Narr** for their input and contribution to the project. I am particularly thankful to **Nils Mechtel,** who teamed up for the software development part. Thank you for your great ideas, diligence, expertise, and enormous work. It was great to work with all of you

on this project. I would also like to acknowledge Biocrates, and especially **Stefan Ledinger** for his support of the project idea and scientific expertise.

Moreover, I am very grateful to **Nicola Zamboni** and his lab. Thank you for welcoming me so warmly in Zurich and making my research stay such a great experience. Thank you for taking all the time to share your expertise in metabolomics with me and the scientific discussion with you.

I also would like to thank my students and interns **Eva Holtkamp**, **Kamil Moskal**, and **Ferdinand Popp** for your substantial contributions, ideas, and enthusiasm for the projects. I could learn a lot from supervising you.

Further, I would like to thank **Sabine Jung-Klawitter** for your collaboration, former supervision, and great ideas throughout the years.

A big thank goes to all present and former members of HI-STEM and A010. Thank you for giving me such a warm welcome from the start. You make this a unique place and I have enjoyed working with all of you.

I am especially thankful to have met wonderful colleagues and friends. Thank you for all the memorable moments during lunch breaks, coffee breaks, dinners, parties, conferences, and holidays. You are all amazing people, and I really enjoyed my time with you from day one. Many thanks to the lunch group, especially **Aino-Majia Leppä**, **Andreas Narr**, **Jonas Schwickert**, **Sarah-Jane Neuberth**, **Simon Renders,** and **Vera Thiel**. Thank you, **Lea Jopp-Saile** and **Florian Grünschläger**, you were the best roommates. Thanks for all the troubleshooting and coding support.

While being everywhere and nowhere, I am particularly grateful for all the collaboration and for always feeling like part of the team. Thank you to all AT club members for your scientific ideas during meetings and your support. Thank you to the Haas Lab, for the meetings during the lockdown and scientific discussions. Thank you, **Simon Haas**, for your enthusiasm for science, your expertise, and great ideas. Thank you, **Megan Druce**, **Marleen Büchler-Schäff**, and **Jayan Jayarajan** for welcoming me in your office. I really enjoyed my time in my office place number two.

In addition, I would like to thank **Dagmar Wolf** and **Erika Krückel**, because of you the bureaucracy became a joy. Thank you for the great administrative support and always being helpful, even last-minute.

I would also like to thank the cluster admins, especially **Frank Thommen** and **Martin Lang**, and the DKFZ Omics IT and Data Management Core Facility (ODCF) for their help, technical support, and for providing such a great infrastructure.

I would also like to acknowledge the **Joachim Herz Stiftung** for supporting my research and especially thank **Karin Liau** for coordinating fellow meetings and her support during the past years.

Thank you to my friends inside and outside of Heidelberg, for freeing my mind from science but also being fellow sufferers on my PhD journey.

I especially thank my parents **Claudia** and **Jens**, my sisters **Annika** and **Melina**, and my family for your support throughout my life. Thank you for always being there for me, encouraging me, believing in me, and giving me the freedom to pursue my way.

Last but not least, I thank **Arne** for your love and support since the early days of my studies. Thank you for being my partner in all our adventures and always being there for me. Without you, I would not have made it!

7 Own publications

First or shared first authorship

Andresen, C.*, Boch, T.*, Hagen, M.G., Mechtel, N., Narr, A., Birgin, E., Rasbach, E., Rahbari, N., Trumpp, A., Poschet, G. Hübschmann, D. (2022). Comparison of extraction methods for intracellular metabolomics of human tissues. Front. Mol. Biosci. 9. https://doi.org/10.3389/fmolb.2022.932261

Hübschmann, D.*, Jopp-Saile, L.*, **Andresen, C.***, Krämer, S., Gu, Z., Heilig, C.E., Kreutzfeldt, S., Teleanu, V., Fröhling, S., Eils, R., et al. (2021). Analysis of mutational signatures with yet another package for signature analysis. Genes Chromosom. Cancer 60, 314–331. https://doi.org/10.1002/gcc.22918

Submitted manuscripts with shared first authorship

Donato, E.*, Correia, N.*, **Andresen, C.***, Karpova, D., Würth, R., Klein, C., Sohn, M., Przybylla, A., Rothfelder, K., Salih, et al. (2022) Retained functional normal and pre-leukemic HSCs at diagnosis are associated to good prognosis in DNMT3A^{mut}NPM1^{mut} AMLs. *In revision at Blood Advances (05.08.2022)*

Other authorship

Heilig, C.E., Laßmann, A., Mughal, S.S., Mock, A., Pirmann, S., Teleanu, V., Renner, M., **Andresen, C.**, Köhler, B.C., Aybey, B., et al. (2022). Gene expression-based prediction of pazopanib efficacy in sarcoma. Eur. J. Cancer 172, 107–118. https://doi.org/10.1016/j.ejca.2022.05.025

Hernández-Malmierca, P., Vonficht, D., Schnell, A., Uckelmann, H.J., Bollhagen, A., Mahmoud, M.A.A., Landua, S.L., van der Salm, E., Trautmann, C.L., Raffel, S., …, **Andresen, C.**, et al. (2022). Antigen presentation safeguards the integrity of the hematopoietic stem cell pool. Cell Stem Cell 29, 760-775.e10. https://doi.org/10.1016/j.stem.2022.04.007

Gegner, H.M., Mechtel, N., Heidenreich, E., Wirth, A., Cortizo, F.G., Bennewitz, K., Fleming, T., **Andresen, C.**, Freichel, M., Teleman, A.A., et al. (2022). Deep Metabolic Profiling Assessment of Tissue Extraction Protocols for Three Model Organisms. Front. Chem. 10. https://doi.org/10.3389/fchem.2022.869732

Jayavelu, A.K., Wolf, S., Buettner, F., Alexe, G., Häupl, B., Comoglio, F., Schneider, C., Doebele, C., Fuhrmann, D.C., Wagner, S., …, **Andresen, C.**, et al. (2022). The proteogenomic subtypes of acute myeloid leukemia. Cancer Cell 40, 301-317.e12. https://doi.org/10.1016/j.ccell.2022.02.006

Sieber-Frank, J., Stark, H.J., Kalteis, S., Prigge, E.S., Köhler, R., **Andresen, C.**, Henkel, T., Casari, G., Schubert, T., Fischl, W., et al. (2021). Treatment resistance analysis reveals GLUT-1-mediated glucose uptake as a major target of synthetic rocaglates in cancer cells. Cancer Med. 10, 6807–6822. https://doi.org/ 10.1002/cam4.4212

Sun, R., He, L., Lee, H., Glinka, A., **Andresen, C.**, Hübschmann, D., Jeremias, I., Müller-Decker, K., Pabst, C., and Niehrs, C. (2021). RSPO2 inhibits BMP signaling to promote self-renewal in acute myeloid leukemia. Cell Rep. 36. https://doi.org/10.1016/j.celrep.2021.109559

Mosqueira, M., Konietzny, R., **Andresen, C.**, Wang, C., and H.A. Fink, R. (2021). Cardiomyocyte depolarization triggers NOS-dependent NO transient after calcium release, reducing the subsequent calcium transient. Basic Res. Cardiol. 116, 1–21. https://doi.org/10.1007/s00395-021-00860-0

Quintero, A., Hübschmann, D., Kurzawa, N., Steinhauser, S., Rentzsch, P., Krämer, S., **Andresen, C.**, Park, J., Eils, R., Schlesner, M., et al. (2021). ShinyButchR: Interactive NMF-based decomposition workflow of genome-scale datasets. Biol. Methods Protoc. 5, 1–7. https://doi.org/10.1093/biomethods/bpaa022

Hartl, C.A., Bertschi, A., Puerto, R.B., **Andresen, C.**, Cheney, E.M., Mittendorf, E.A., Guerriero, J.L., and Goldberg, M.S. (2019). Combination therapy targeting both innate and adaptive immunity improves survival in a pre-clinical model of ovarian cancer. J. Immunother. Cancer 7, 1–17.
https://doi.org/10.1186/s40425-019-0654-5

Software

Mechtel, N.*, **Andresen, C.***, and Hübschmann, D. (2022). MetAlyzer: Read and Analyze "MetIDQ" Software Output Files. 0.1.0. https://CRAN.R-project.org/package=MetAlyzer

R Shiny app MetaboExtract: https://www.metaboextract.shiny.dkfz.de/MetaboExtract/

* shared first authors

8 Appendix

8.1 Supplementary Figures

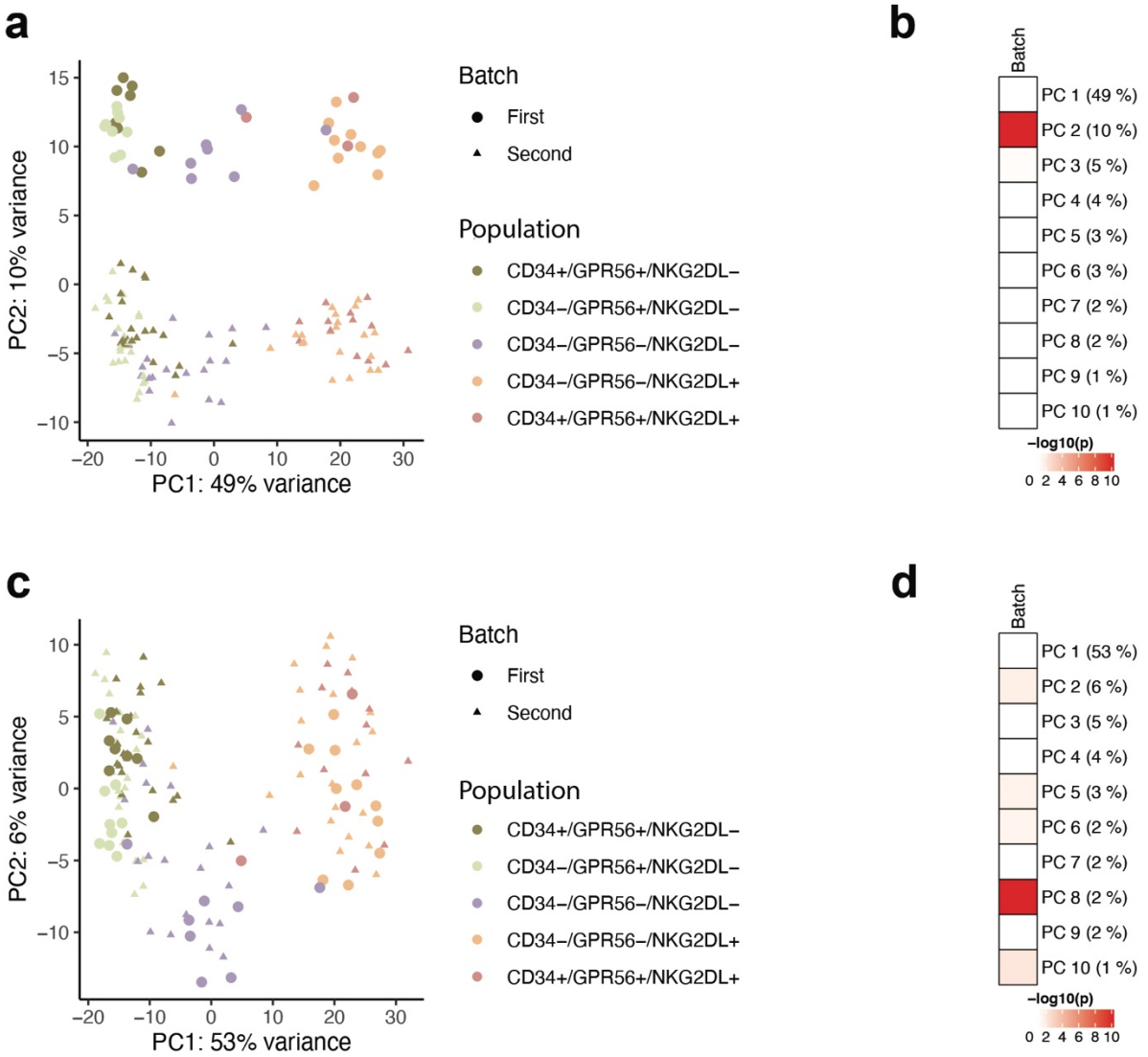

Figure S 1: Correction of two RNA-seq batches. a) PCA before batch correction. b) Kruskal-Wallis Rank Sum Test between batch information and first 10 principal components. c) PCA after batch correction. d) Kruskal-Wallis Rank Sum Test between batch information and first 10 principal components.

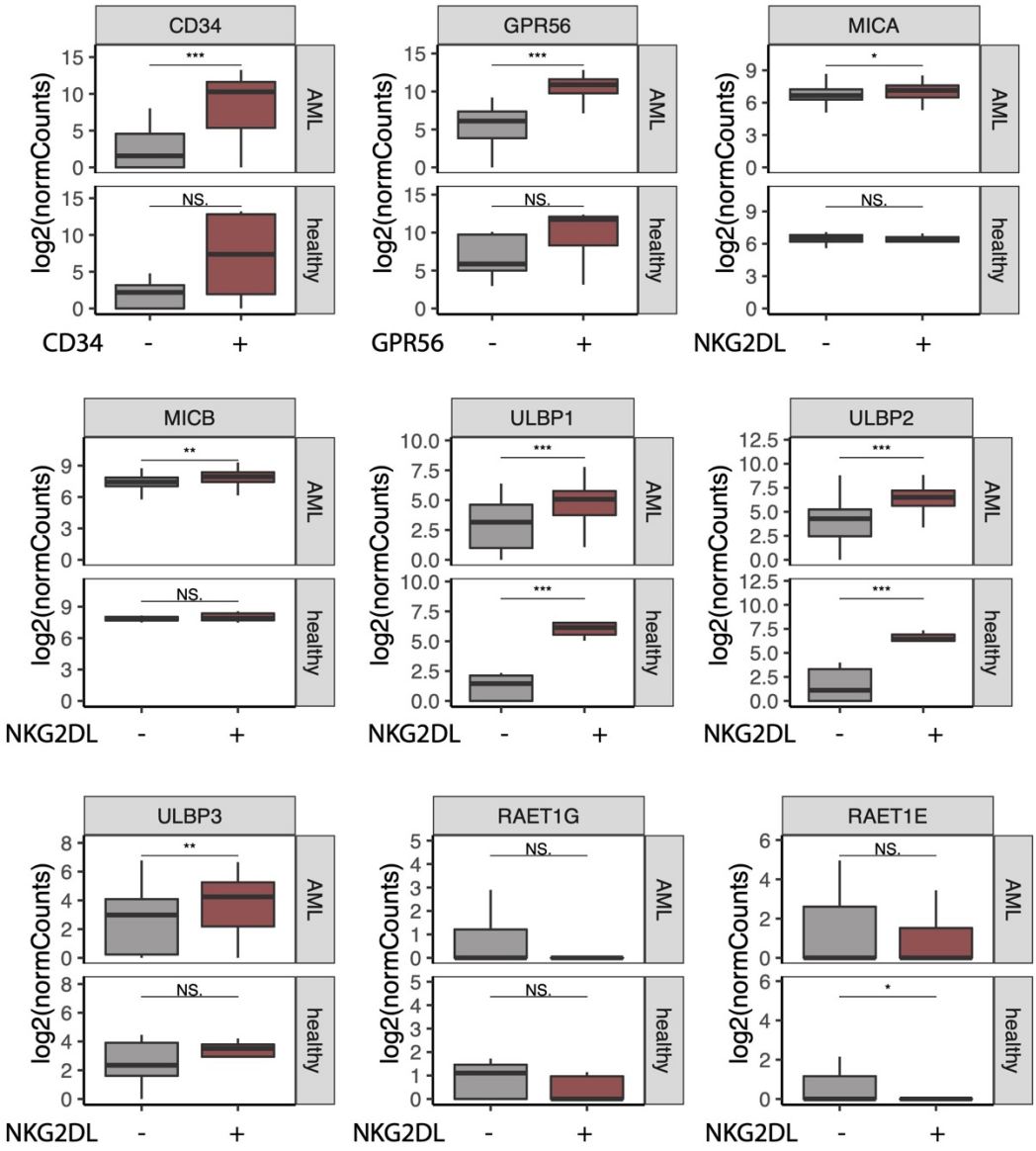

Figure S 2: Expression of marker genes in FACS negative and positive cells. NKG2D ligand *RAER1L* expression was not detected. *RAET1E* and *RAET1G* are also known as *ULBP4* and *ULBP5*, respectively. Statistical differences were calculated between groups using a two-sided Wilcoxon Rank Sum Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

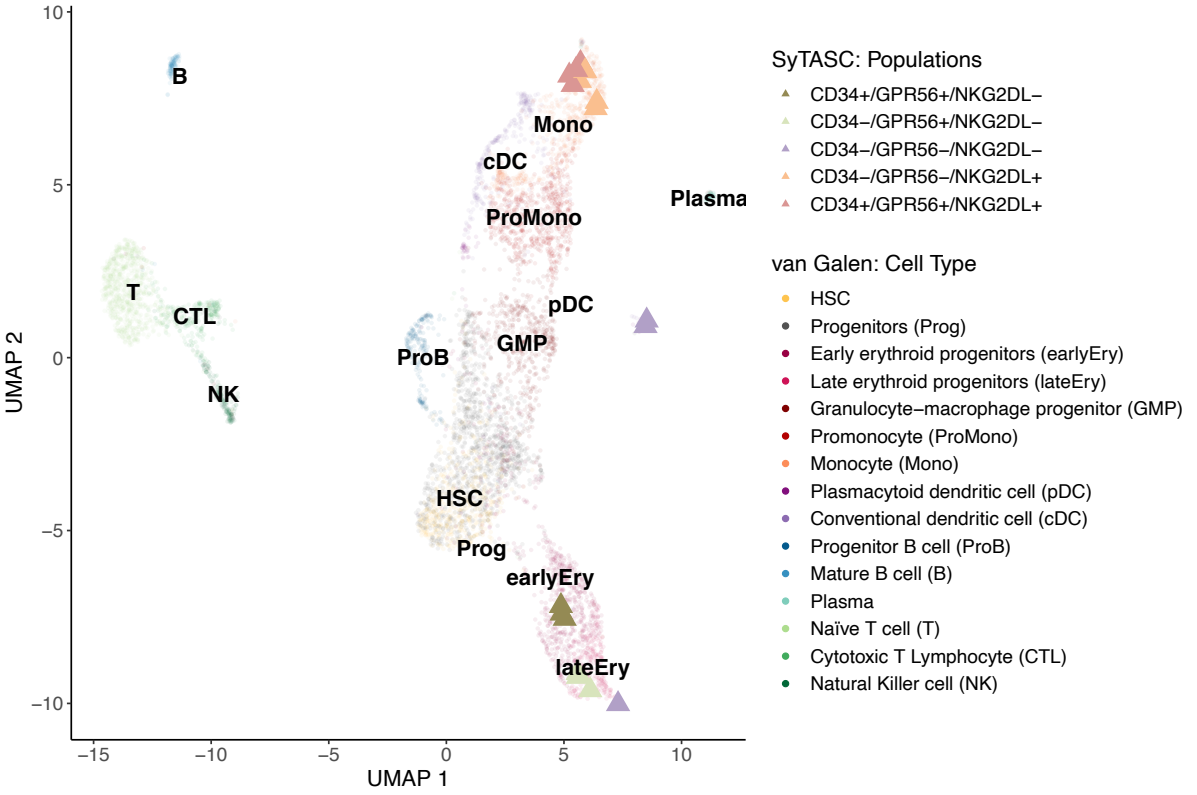

Figure S 3: Embedding of healthy sorted populations into UMAP of single-cell data published by van Galen et al. as reference.[40]

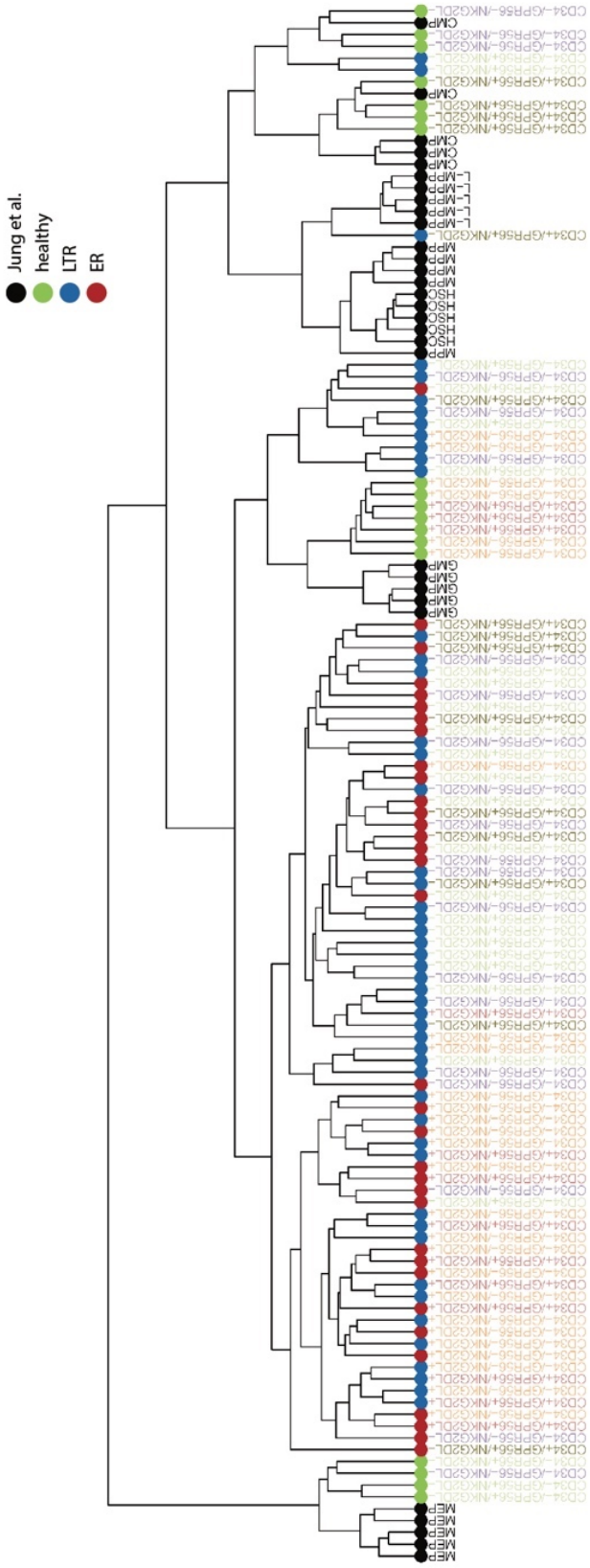

Figure S 4: Hierarchical clustering of 216 differentially methylated regions in healthy HSPCs compared with sorted populations. MEP: megakaryocytic-erythroid progenitor; GMP: granulocyte-macrophage progenitor; MPP: multipotent progenitor; HSC: hematopoietic stem cell; L-MPP: lymphoid-primed multipotential progenitor; CMP: common myeloid progenitor.[196]

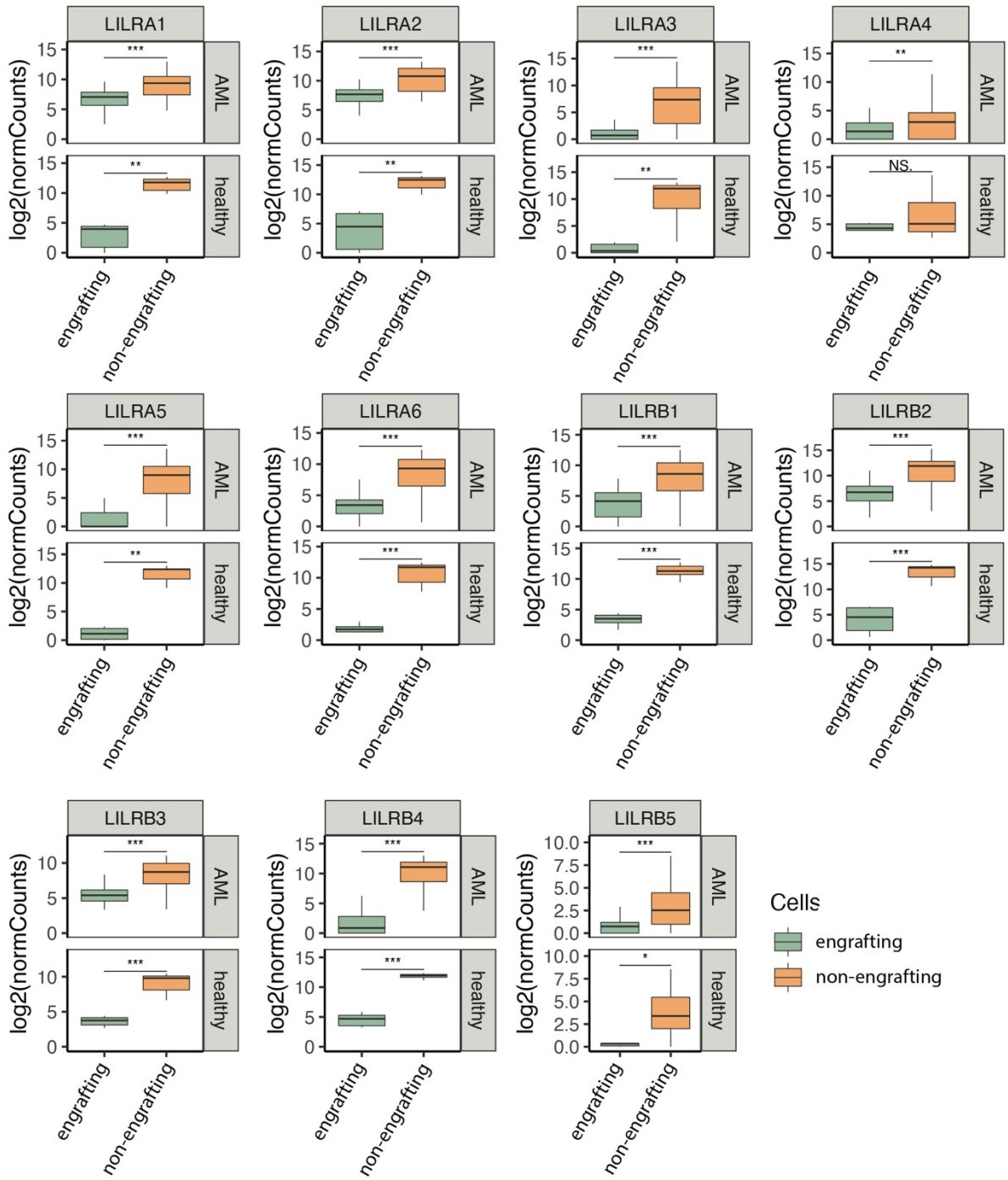

Figure S 5: Expression of LILRs in engrafting compared to non-engrafting populations. Statistical differences were calculated between groups using a two-sided Wilcoxon Rank Sum Test: *p ≤ 0.05; **p ≤ 0.01; *** p ≤ 0.001; NS: non-significant.

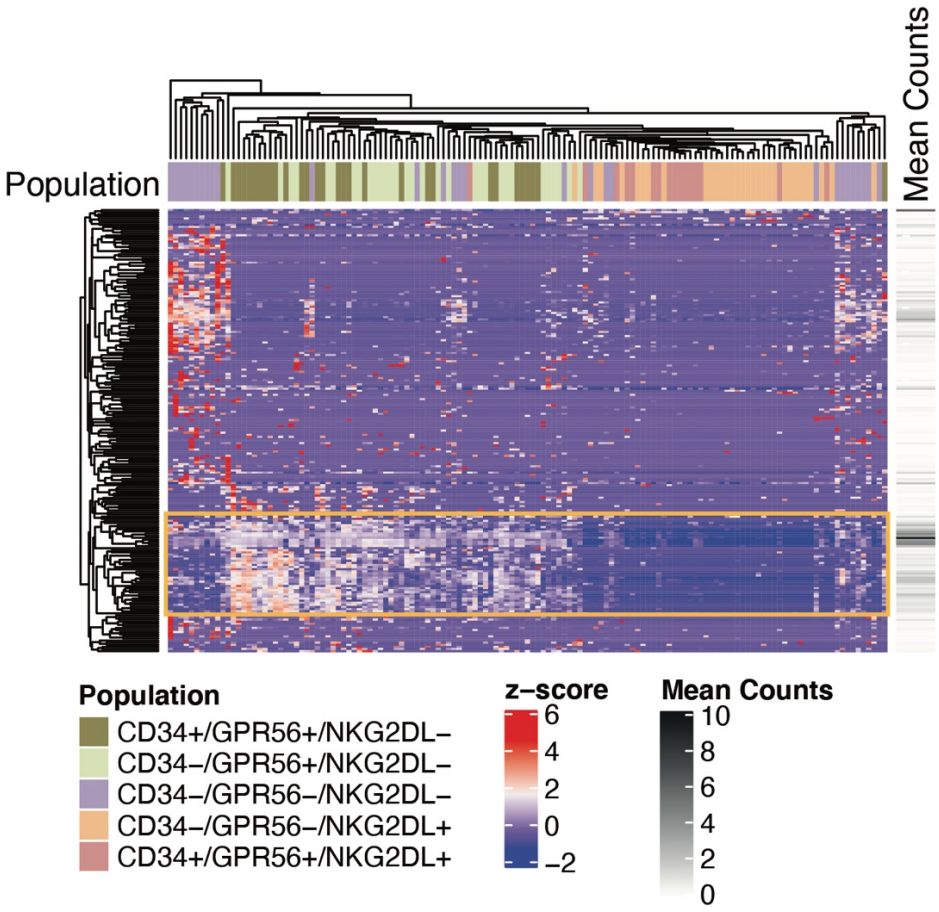

Figure S 6: Heatmap showing expression of immunoglobulin genes. The yellow box highlights genes that are clearly differential between sorted populations.

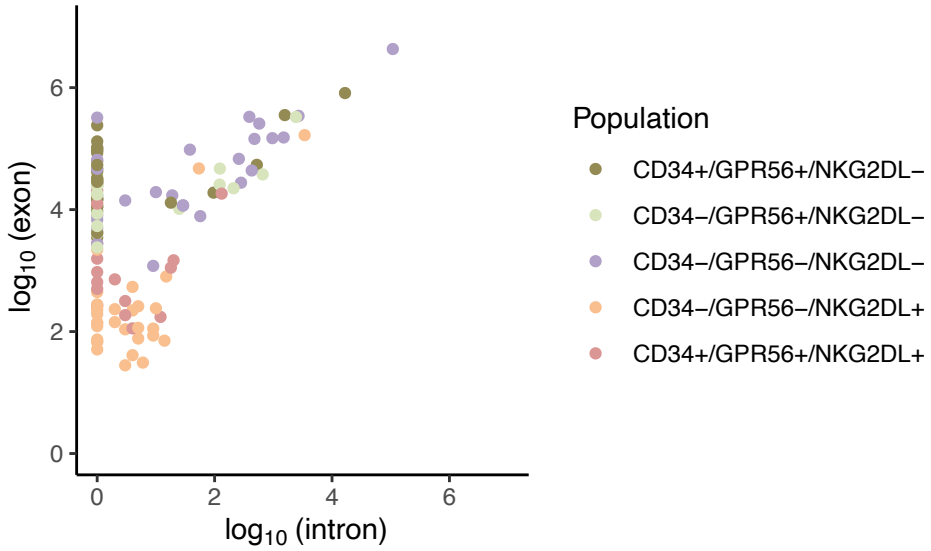

Figure S 7: Intronic and exonic count analysis of immunoglobulin genes. The Scatter plot shows the sum of normalized intro- exon counts per sample. In total, 360 immunoglobulins could be quantified since only non-overlapping genes can be analyzed. Sum of exon counts: 10,379,021; sum of intron counts 142,581. 91% of intron counts can be assigned to one sample (SyT 74).

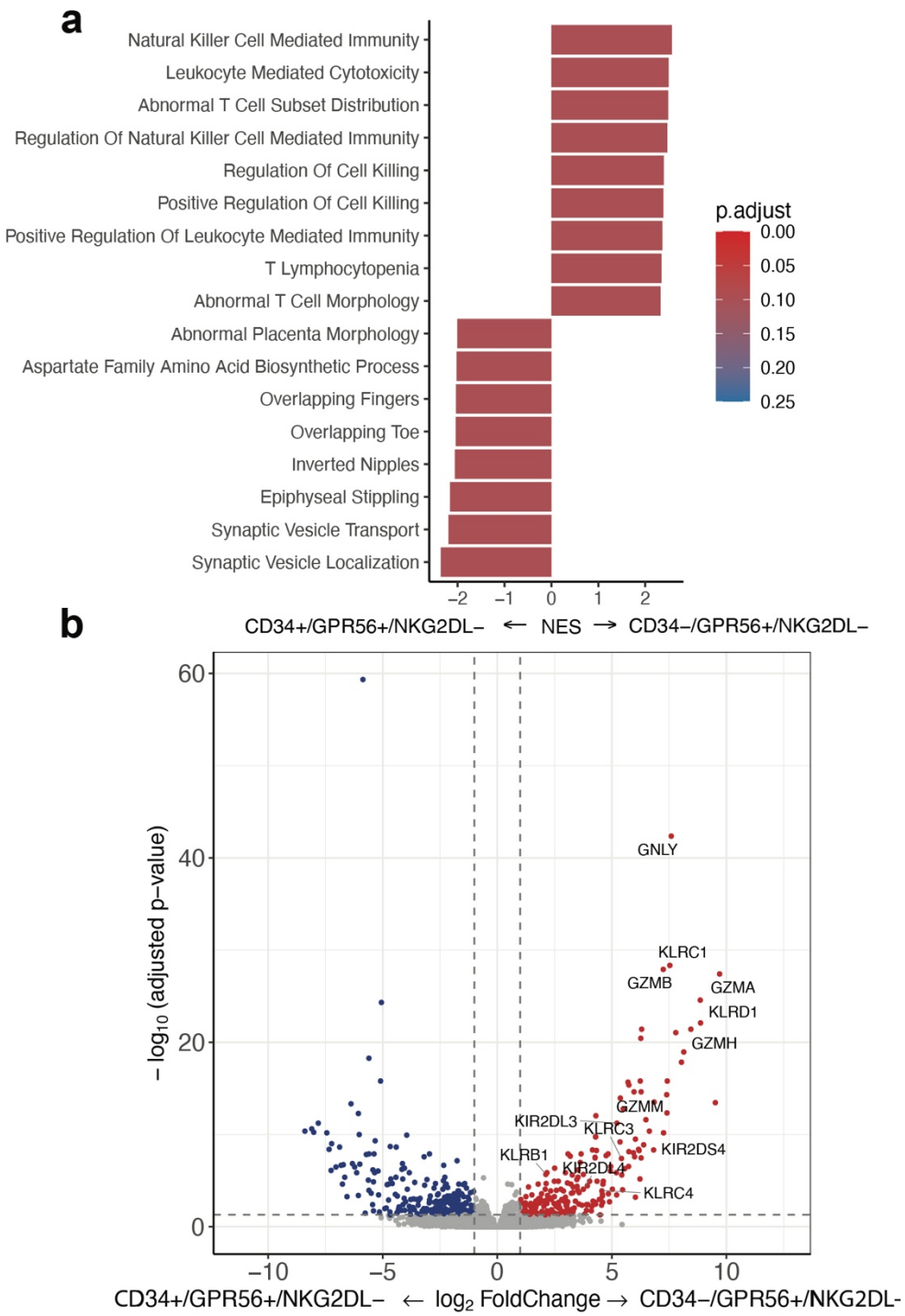

Figure S 8: Differential expression and GSEA between CD34^{+}GPR56^{+}NKG2DL^{-} and CD34^{-}GPR56^{+}NKG2DL^{-} LSC populations in the "original" data set. a) GSEA of most significant gene sets. b) Volcano plot highlighting NK cell-related expression.

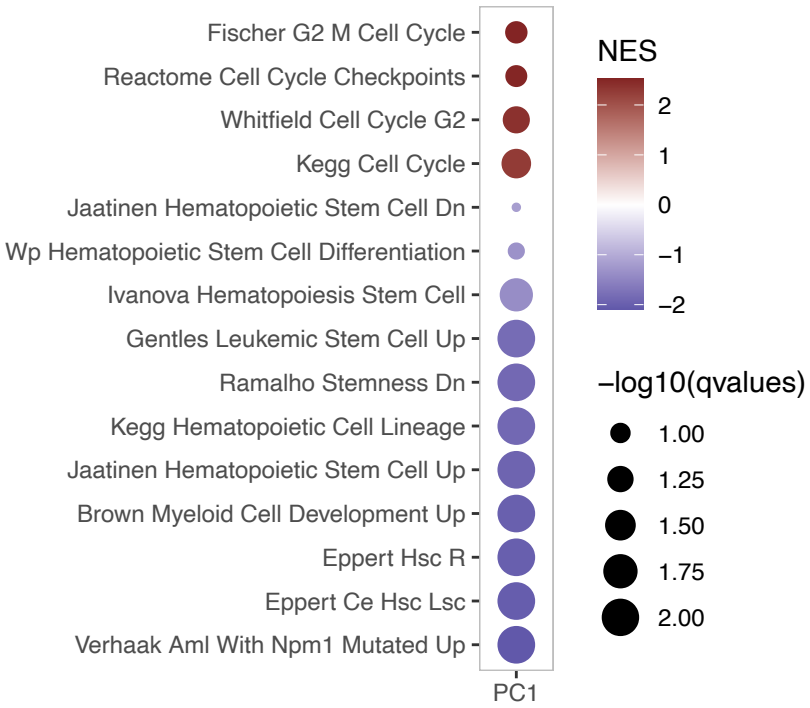

Figure S 9: Selected gene sets from GSEA of PC1 shown in Figure 19a.

Table S 1: Engraftment of samples for bulk and sorted populations. BMT: bone marrow transplantation; m.-lin.: multi-lineage. Samples with grey shading only bulk sequencing was performed. For sample SYT84* bulk sequencing was not performed. NA indicates that sample was not sequenced or transplanted, e.g., because of low cell numbers.

Patient	Outcome Group	Bulk Engraftment	$CD34^+$ $GPR56^+$ $NKG2DL^-$	$CD34^-$ $GPR56^+$ $NKG2DL^-$	$CD34^-$ $GPR56^-$ $NKG2DL^-$	$CD34^-$ $GPR56^-$ $NKG2DL^+$	$CD34^+$ $GPR56^+$ $NKG2DL^+$
SYT03	ER	AML	AML	AML	No	No	No BMT
SYT06	ER	AML	NA	NA	NA	NA	NA
SYT08	ER	AML	m.-lin.	No	No	No	NA
SYT17	ER	AML	No	No	No	No	No
SYT18	ER	AML	AML	AML	No	No	NA
SYT20	ER	AML	AML	AML	No	No	NA
SYT22	ER	AML	AML	AML	No	No	No
SYT25	ER	AML	AML	AML	No	No	Yes [?]
SYT27	ER	AML	AML	No	No	No	No
SYT30	ER	AML	AML	No	No	No	NA
SYT34	ER	AML	AML	No	No	No	No BMT
SYT39	ER	AML	No	Yes [?]	No	No	NA
SYT01	LTR	AML	m.-lin.	AML	No	No	No
SYT02	LTR	AML	AML	AML	No	No	NA
SYT07	LTR	AML	No	No	NA	No	AML
SYT10	LTR	m.-lin.	NA	No	No	No	NA
SYT14	LTR	m.-lin.	NA	NA	NA	NA	NA
SYT23	LTR	m.-lin.	No	No	No	No	No
SYT26	LTR	AML	NA	NA	NA	NA	NA
SYT29	LTR	AML	m.-lin.	No	No	No	No
SYT32	LTR	No	NA	NA	NA	NA	NA
SYT37	LTR	No	NA	NA	NA	NA	NA
SYT42	LTR	m.-lin.	No BMT	No BMT	No BMT	No BMT	No BMT
SYT45	LTR	AML	No	No	No	No	NA
SYT47	LTR	AML	AML	AML	No	No	NA
SYT50	LTR	m.-lin.	m.-lin.	AML	No	No	NA
SYT51	LTR	AML	AML	No	No	No	NA
SYT52	LTR	m.-lin.	NA	NA	NA	NA	NA
SYT56	LTR	AML	No	AML	NA	No	NA
SYT59	LTR	AML	AML	AML	No	No	NA
SYT61	LTR	unclear	NA	NA	NA	NA	NA
SYT64	LTR	AML	AML	AML	No	No	No
SYT67	LTR	AML	No	No	No	No	No BMT
SYT74	LTR	AML	No	No	No	No	NA
SYT76	LTR	m.-lin.	No	No	No	No	No
SYT84*	LTR	m.-lin.	No BMT	No BMT	No BMT	No BMT	NA
SYT88	LTR	m.-lin.	No	No	No	No	No BMT
SYT90	LTR	m.-lin.	AML	No	No	No	No

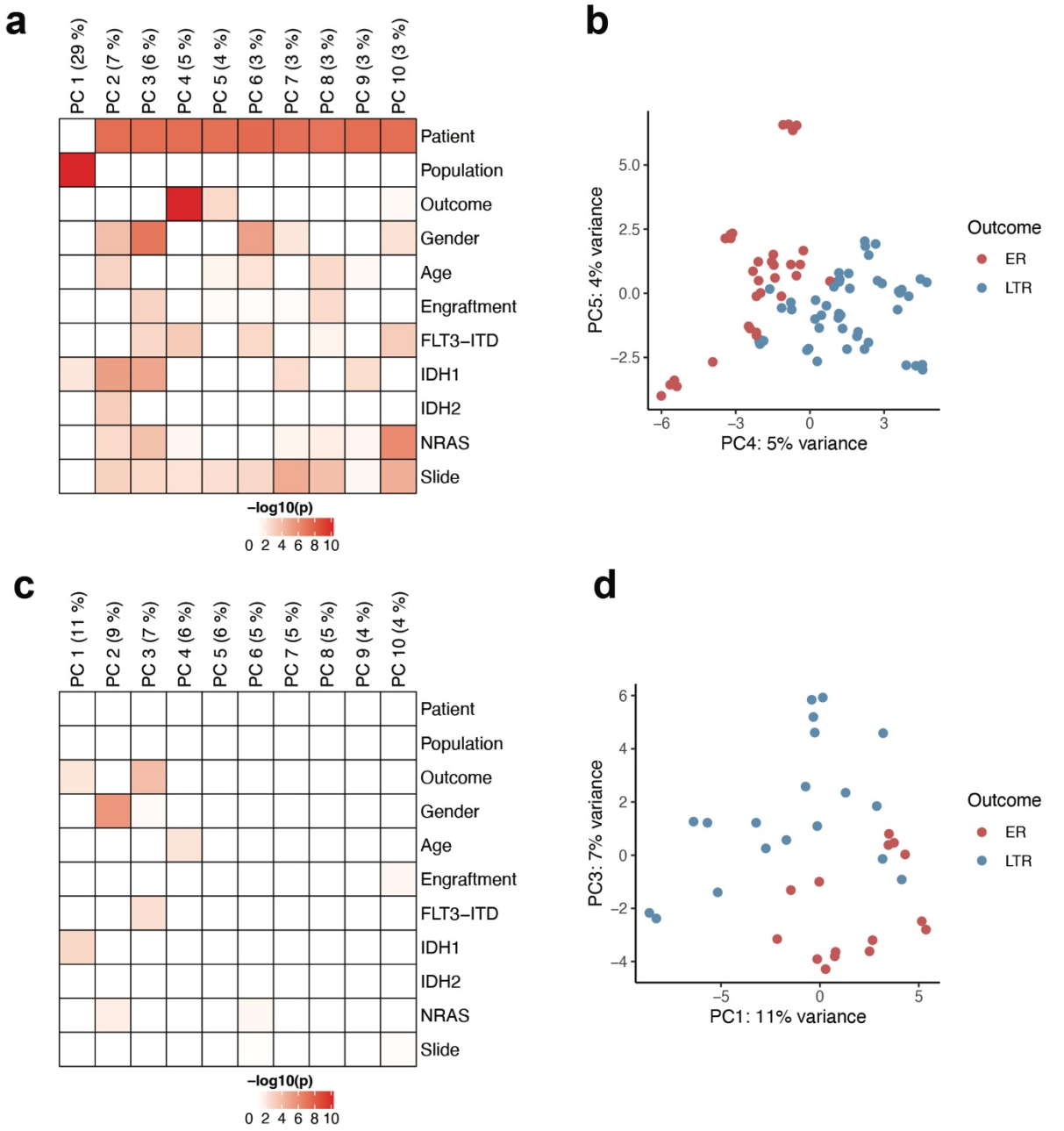

Figure S 10: Sources of variability in methylation data. a,c) Heatmap showing p-values of Kruskal-Wallis Rank Sum Test between biological or clinical information and first 10 principal components for: a) all populations. c) Only engrafting LSC populations. b,d) PCA for: b) all populations. d) Only engrafting LSC populations.

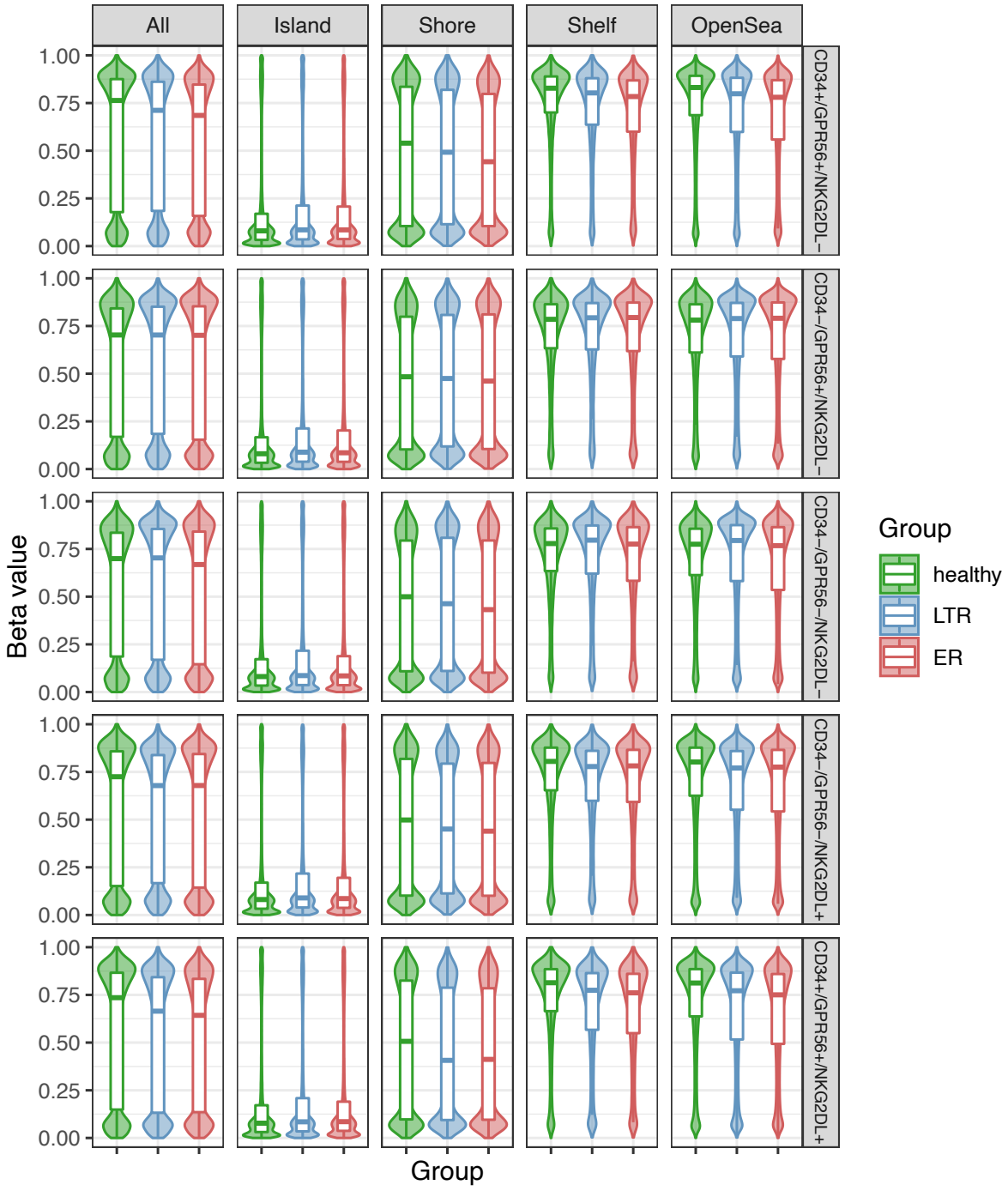

Figure S 11: Distribution of beta values for all regions, island, shore, shelf, and open sea and for all sorted populations between healthy, LTR, and ER samples.

Table S 2: Quantification of alternative splicing events corresponding to the violin plots presented in Figure 33. RI: intron retention, MXE: mutually exclusive exons, SE: exon skipping, A5SS: alternative 5' donor site, A3SS: alternative 3' donor site. Statistics was filtered for event with an absolute inclusion level difference > 0.1 and an FDR < 0.05.

	LSC-enriched populations		**All populations**	
Event Types	ΔIncLevel > 0	ΔIncLevel < 0	ΔIncLevel > 0	ΔIncLevel < 0
A3SS	220	237	138	179
A5SS	192	142	143	123
MXE	1,017	871	1,233	855
RI	306	103	172	62
SE	1,277	1,164	738	1,168

Table S 3: Comparison of differentially spliced and methylated genes. Significant alternative splicing was defined by an absolute inclusion level difference > 0.1 and an FDR < 0.05. Differential methylated regions were considered significant for an absolute beta value difference > 0.25 and FWER < 0.05. Fisher's Exact Test on numbers of genes estimated a p-value = 0.019.

Event Types		**Methylation**	
	Significance	Differential	Not differential
Alternative Splicing	Differential	11	1,889
	Not Differential	145	55,620

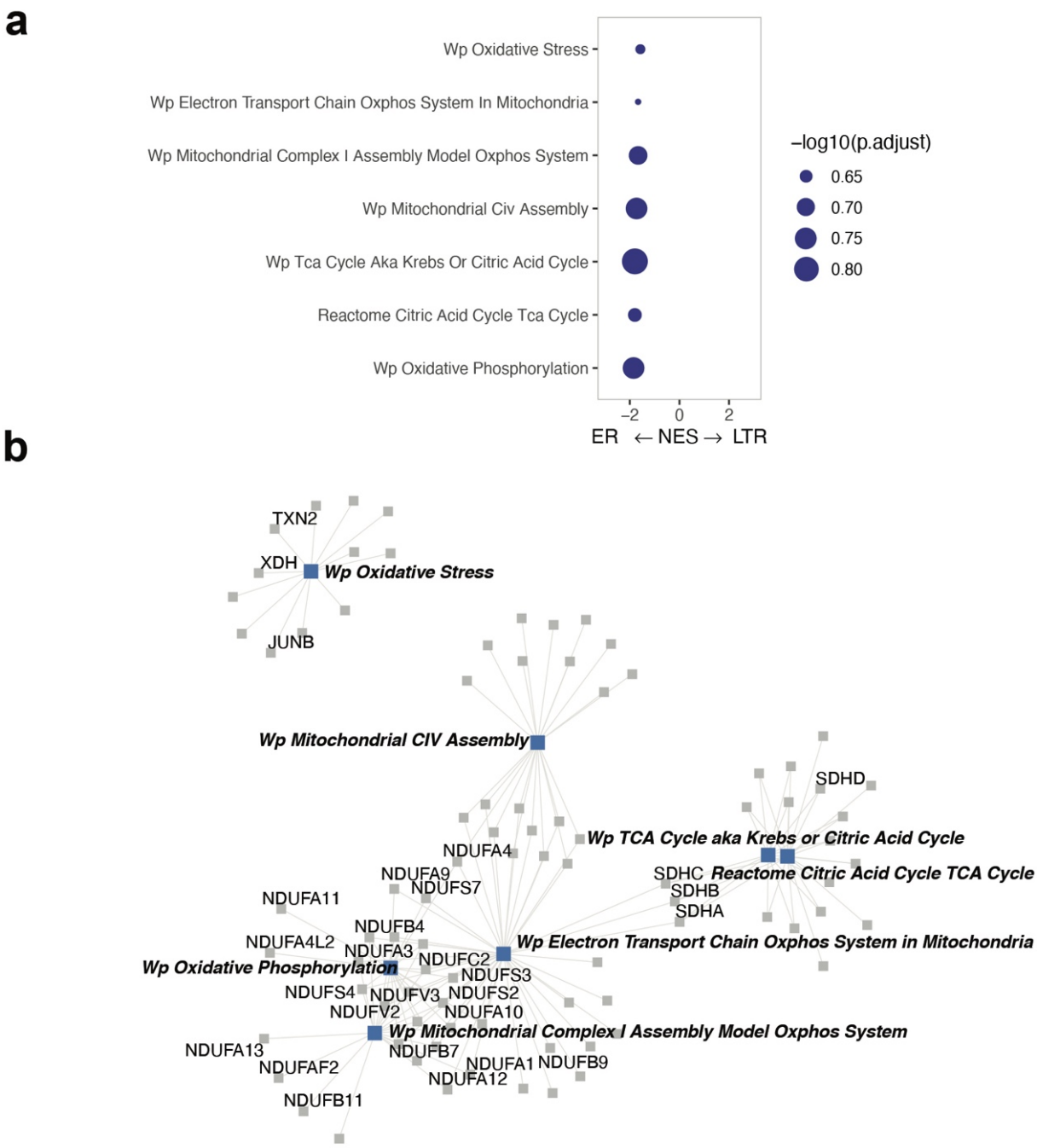

Figure S 12: Enrichment of gene sets and pathways related to oxidative phosphorylation and mitochondrial complexes in the bulk data set. a) Bar plot showing selected GSEA gene sets between LTR and ER. b) Selected genes identified by leading-edge analysis.

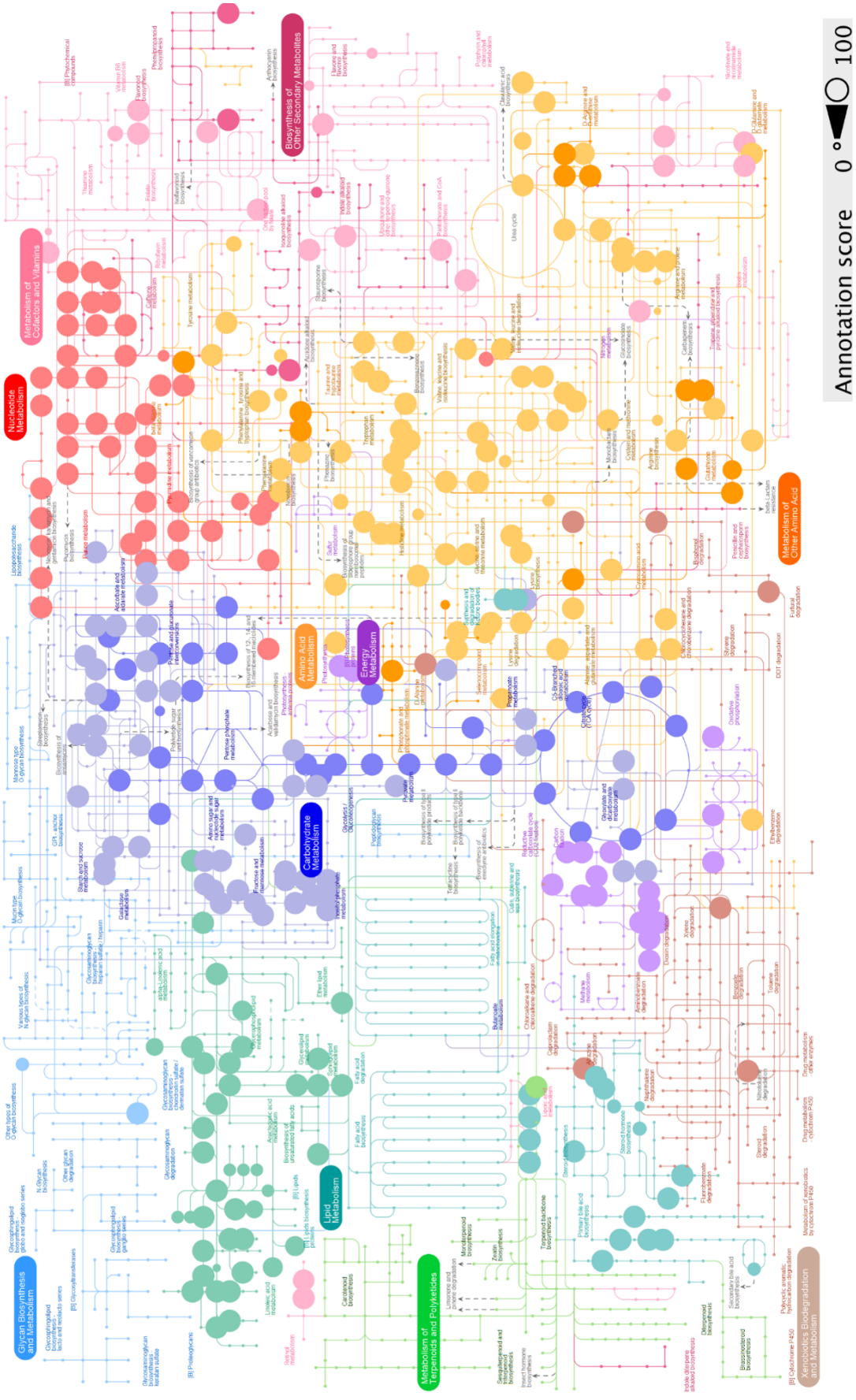

Figure S 13: Coverage of untargeted metabolomics measurements.

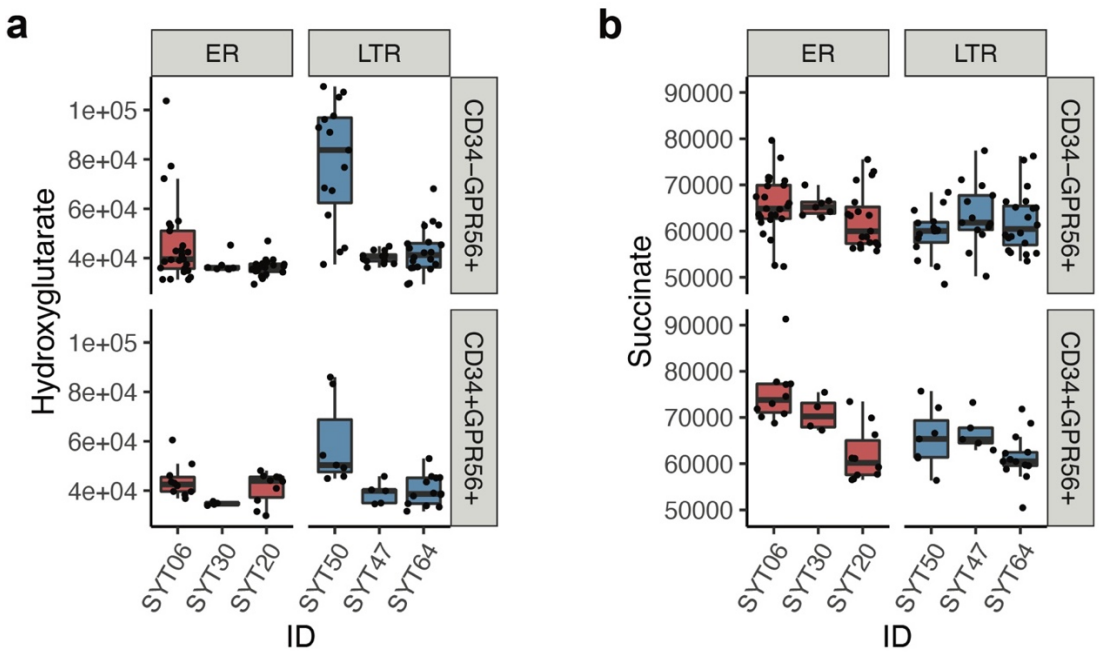

Figure S 14: Abundance of hydroxyglutarate and succinate. SYT50 is an IDH1-mutant AML, while the other samples were IDH1-wt.

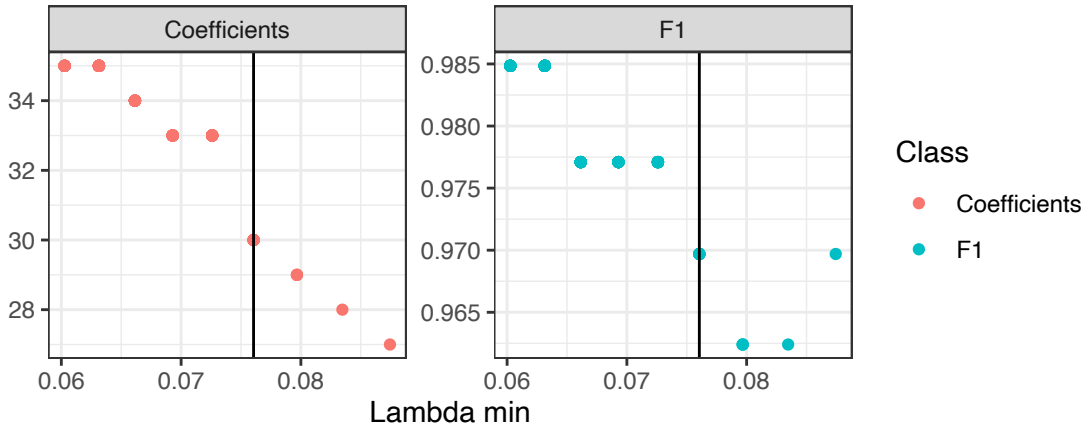

Figure S 15: Distribution of the number of coefficients and F1 scores for lambda min values calculated in 100 iterative training runs. The optimal lambda was chosen to be 0.0761.

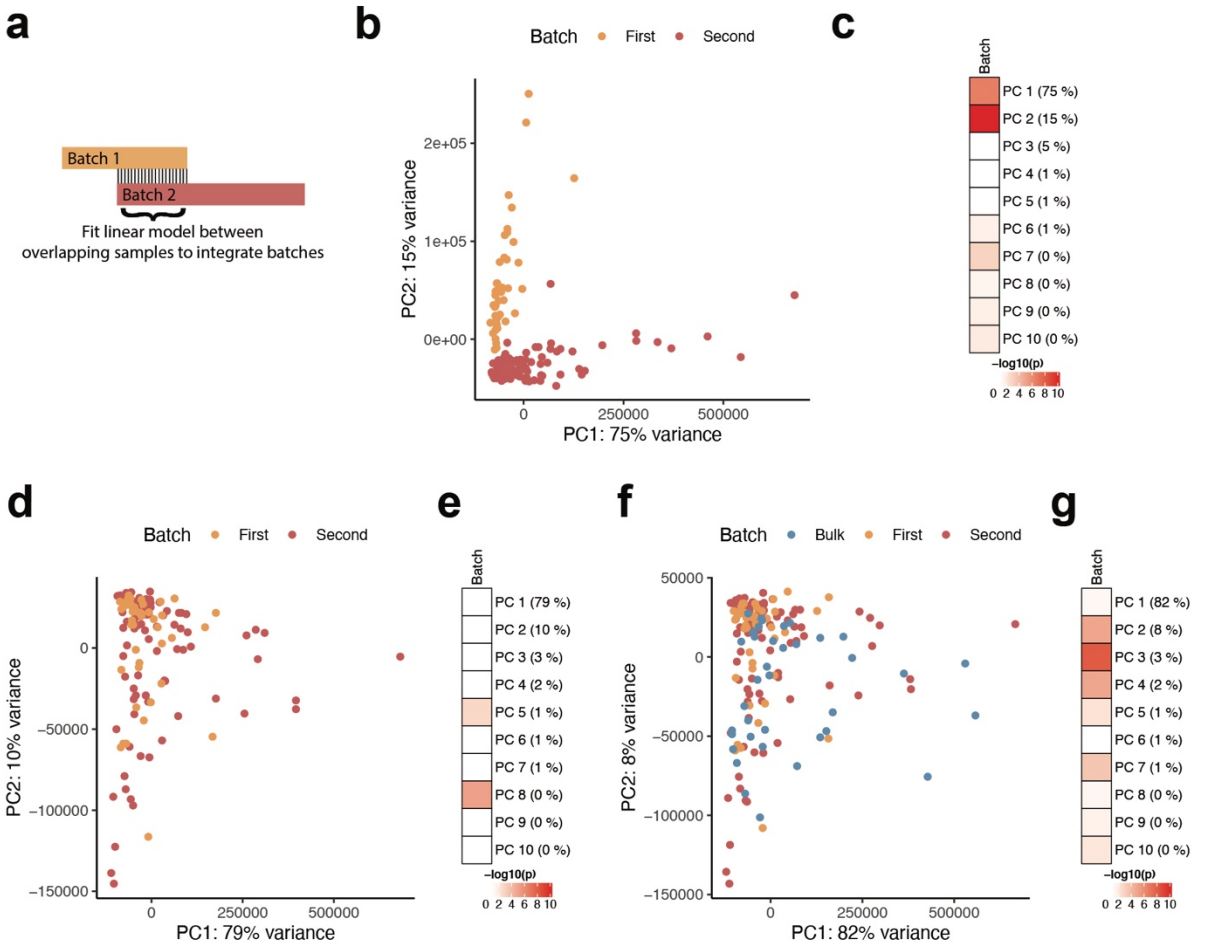

Figure S 16: Customized batch correction of TPMs for the training of outcome prediction signature. a) Overview approach. b-g) PCA and heatmap showing p-values of Kruskal-Wallis Rank Sum Test between batch information and first 10 principal components for: b-c) Population-sorted data set before batch correction. d-e) Population-sorted data set after batch correction. f-g) Population-sorted data set and bulk data set after batch correction.

Table S 4: Coefficients of signature genes.

Gene	Coefficient	Ensembl	Gene Type
CAMK1D	0.2011423	ENSG00000183049.8	protein_coding
SORT1	0.1632886	ENSG00000134243.7	protein_coding
HLA-DOA	0.09528722	ENSG00000204252.8	protein_coding
BTG2	0.09223349	ENSG00000159388.5	protein_coding
C10orf128	0.08237871	ENSG00000204161.9	protein_coding
ENPP2	0.05360229	ENSG00000136960.8	protein_coding
MPZL2	0.04008115	ENSG00000149573.4	protein_coding
PAX8	0.03936341	ENSG00000125618.12	protein_coding
MRC1L1	0.03413372	ENSG00000183748.4	protein_coding
S100B	0.01933282	ENSG00000160307.5	protein_coding
AC061992.1	-0.0005678	ENSG00000268965.1	protein_coding
PLD1	-0.0008052	ENSG00000075651.11	protein_coding
RHD	-0.0136507	ENSG00000187010.14	protein_coding
TRIP6	-0.0238733	ENSG00000087077.7	protein_coding
NDRG1	-0.0351277	ENSG00000104419.10	protein_coding
SEMA4B	-0.0377146	ENSG00000185033.10	protein_coding
SLITRK4	-0.0436313	ENSG00000179542.11	protein_coding
CLDN10	-0.0445897	ENSG00000134873.5	protein_coding
CTH	-0.0683028	ENSG00000116761.7	protein_coding
RFX2	-0.0785434	ENSG00000087903.8	protein_coding
TMEM99	-0.0788538	ENSG00000167920.4	protein_coding
FGFR1	-0.0963102	ENSG00000077782.15	protein_coding
CCL3L3	-0.1050339	ENSG00000256515.3	protein_coding
NPIPA5	-0.1103931	ENSG00000183793.9	protein_coding
ZNF672	-0.1264443	ENSG00000171161.8	protein_coding
EIF2S3L	-0.1607959	ENSG00000180574.3	protein_coding
AZI2	-0.1674812	ENSG00000163512.9	protein_coding
TBX6	-0.1676564	ENSG00000149922.6	protein_coding
FAM131A	-0.2143174	ENSG00000175182.9	protein_coding
CLGN	-0.2674335	ENSG00000153132.8	protein_coding

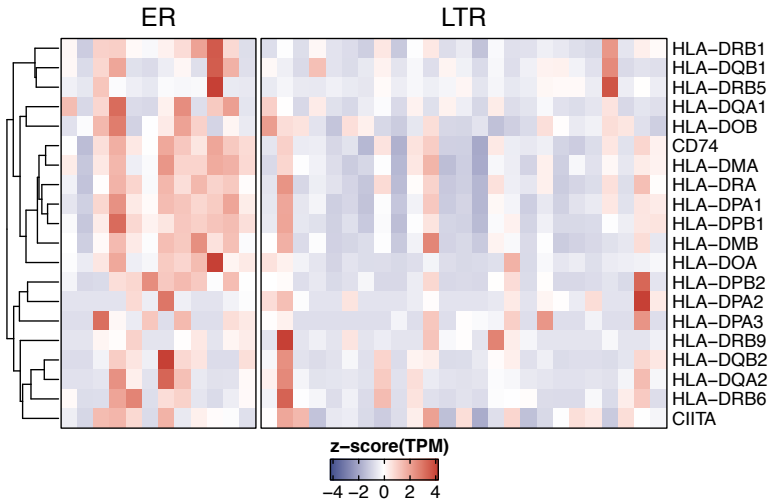

Figure S 17: Heatmap showing expression of MHC-II genes.

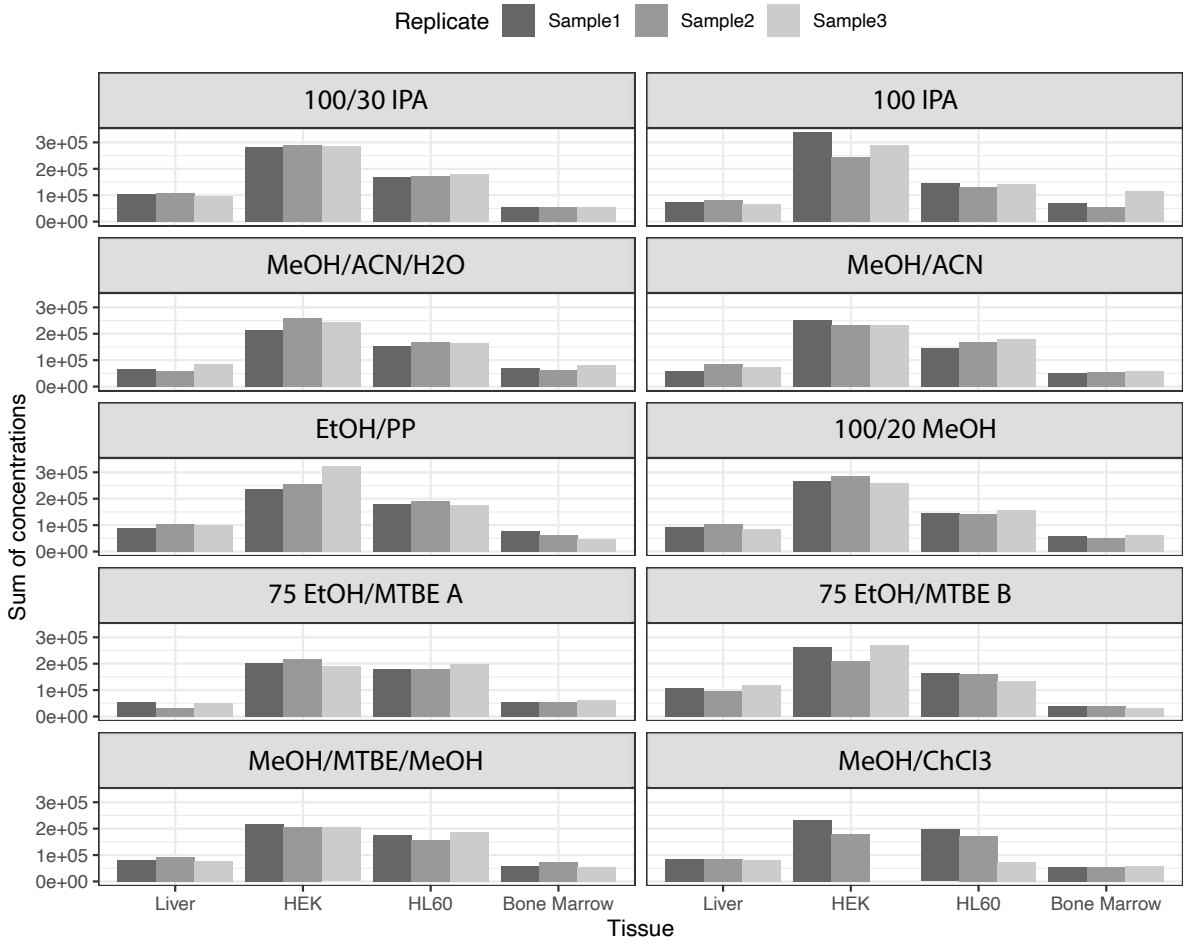

Figure S 18: SOC between replicates across extraction protocols and sample types. Of note, one replicate of HEK was removed because of very low concentrations. HEK, HL60, and bone marrow Concentrations are given in picomole per 10^6 cells and picomole per mg for liver tissue.

8.2 List of abbreviations

A3SS	alternative 3’ donor site
A5SS	alternative 5’ donor site
ACK	ammonium-chloride-potassium
ACN	acetonitrile
ADP	adenosine diphosphate
ALL	acute lymphoblastic leukemia
AML	acute myeloid leukemia
ANOVA	analysis of variance
APC	antigen presenting cell
ATP	adenosine triphosphate
BAM	binary alignment map
BCL6	BCL6 Transcription Repressor
BHLHE40	Basic Helix-Loop-Helix Family Member E40
BM	bone marrow
BMP	bone morphogenetic protein
BMT	bone marrow transplantation
bp	base pairs
CD	cluster of differentiation
CH_3	methyl group
$ChCl_3$	chloroform
CHIP	Clonal hematopoiesis of indeterminate potential
CLP	common lymphoid progenitor
CMP	common myeloid progenitor
CMP	common myeloid progenitor
CO_2	carbon dioxide
CSF	cerebrospinal fluid
CTL	cytotxic T lymphocytes
CV	coefficient of variation
DAPI	4’,6-diamidino-2-phenylindole
DKFZ	Deutsches Krebsforschungszentrum
DMEM	Dulbecco’s Modified Eagle’s Medium
DMP	differentially methylated positions

DMR	differentially methylated regions
DNMT3A	DNA (cytosine-5)-Methyltransferase 3A
E2F4	E2F Transcription Factor 4
EGA	European Genome-phenome Archive
EISA	exon-intron split analysis
ELN	European LeukemiaNet
ENPP2	Ectonucleotide Pyrophosphatase/Phosphodiesterase 2
ER	early relapse
EtOH	ethanol
FAB	French-American-British
FACS	fluorescence-activated cell sorting
FCS	fetal calf serum
FDR	false discovery rate
FPKM	Fragments Per Kilobase Million
FSC A	forward scatter area
FWER	family-wise error rate
GMP	granulocyte-macrophage progenitor
GO	gene ontology
GPR56 (*ADGRG1*)	adhesion G Protein-Coupled Receptor 56
GSEA	gene set enrichment analysis
H_2O	water
HEK	human embryonic kidney
HK3	Hexokinase 3
HL60	human leukemia 60
HLA-DOA	Major Histocompatibility Complex, Class II, DO Alpha
HMDB	Human Metabolome Database
hnRNP	heterogeneous nuclear ribonucleoprotein
HSC	hematopoietic stem cell
HSPC	hematopoietic stem and progenitor cells
ID4	Inhibitor of Differentiation 4
IDH1	Isocitrate Dehydrogenase 1
IMDM	Iscove's Modified Dulbecco's Medium
IPA	isopropanol
JAK	Janus kinase

KIR	Killer Cell Immunoglobulin-Like Receptor
LIC	leukemia-initiating cell
LILR	Leukocyte Immunoglobulin-Like Receptor
LMP	lymphoid-primed multipotent progenitor
LMPP	lymphoid-primed multipotential progenitor
LN_2	liquid nitrogen
LOD	limit of detection
LOLA	Locus Overlap Analysis
LSC	leukemic stem cell
LT	latent factor
LTR	long-term remission
MAD	median absolute deviation
MDS	myelodysplastic syndrome
MeOH	methanol
MEP	megakaryocyte-erythroid progenitor
MEP	megakaryocytic-erythroid progenitor
MHC	major histocompatibility complex
MICA	MHC Class I Polypeptide-Related Sequence A
MICB	MHC Class I Polypeptide-Related Sequence B
MNC	mononuclear cells
MOFA	Multi-Omics Factor Analysis
MPP	multipotent progenitor
mRNA	messenger RNA
MTBE	methyl tert-butyl ether
MXE	mutually exclusive exons
NaCl	sodium chloride
NAD^+(H)	nicotinamide adenine dinucleotide (oxidized/reduced)
NES	normalized enrichment score
NGS	next-generation sequencing
NK	natural killer
NKG2D (*KLRK1*)	Killer Cell Lectin Like Receptor K1 (ligands)
NKG2DL	NKG2D ligands
NNLS	non-negative least squares
NOG	Noggin

NPM1	Nucleophosmin
NS	non-significant
NSG	NOD scid gamma
OS	overall survival
PBS	phosphate-buffered saline
PC	principal component
PC	Pyruvate Carboxylase
PCA	principal component analysis
PD-1	Programmed Cell Death Protein 1
PD-L1	Programmed death-ligand 1
PI3K	Phosphoinositide 3-kinase
PKLR	pyruvate kinase L/R
PP	polypropylene
PROGENy	Pathway RespOnsive GENes for activity inference
QC	quality control
RAET	Retinoic Acid Early Transcript
RBP	RNA-binding protein
RFS	relapse-free survival
RI	intron retention
ROS	reactive oxygen species
RPKM	Reads Per Kilobase Million
RT	room temperature
S100	S100 calcium-binding proteins
SD	standard deviation
SE	exon skipping
SF3B1	Splicing Factor 3b Subunit 1
SFRS2	Serine and Arginine-Rich Splicing Factor 2
snRNP	small ribonucleoprotein complex
SOC	sum of concentrations
SOP	standard operating procedure
SORT1	Sortilin 1
SOX2	SRY-Box Transcription Factor 2
SPI1	Spi-1 Proto-Oncogene
SR protein	serine and argine-rich protein

SSC A	side scatter area
STAT	Signal Transducer and Activator of transcription
SyTASC	Systems-based Therapy of AML Stem Cells
TCA	tricarboxylic acid
TGFβ	Transforming Growth Factor β
TLR4	Toll Like Receptor 4
TPM	transcripts per million
ULBP	UL16 Binding Protein
UMAP	Uniform Manifold Approximation and Projection
VAF	variant allele frequencies
VIPER	Virtual Inference of Protein-activity by Enriched Regulon
vst	variance-stabilizing transformation
WBC	white blood cells
WES	whole exome sequencing
WGS	whole genome sequencing
WHO	World Health Organization

8.3 List of figures

8.4 List of tables

9 Bibliography

1. Shallis, R.M., Wang, R., Davidoff, A., Ma, X., and Zeidan, A.M. (2019). Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. Blood Rev. *36*, 70–87.
2. Yanada, M., and Naoe, T. (2012). Acute myeloid leukemia in older adults. Int. J. Hematol. *96*, 186–193.
3. Döhner, H., Weisdorf, D.J., and Bloomfield, C.D. (2015). Acute Myeloid Leukemia. N. Engl. J. Med. *373*, 1136–1152.
4. Newell, L.F., and Cook, R.J. (2021). Advances in acute myeloid leukemia. BMJ *375*, n2026.
5. Bernt, K.M. (2017). Bridging the Gaps: iPSC-Based Models from CHIP to MDS to AML. Cell Stem Cell *20*, 298–299.
6. Laverdière, I., Boileau, M., Neumann, A.L., Frison, H., Mitchell, A., Ng, S.W.K., Wang, J.C.Y., Minden, M.D., and Eppert, K. (2018). Leukemic stem cell signatures identify novel therapeutics targeting acute myeloid leukemia. Blood Cancer J. *8*.
7. Clarke, M.F., Dick, J.E., Dirks, P.B., Eaves, C.J., Jamieson, C.H.M., Jones, D.L., Visvader, J., Weissman, I.L., and Wahl, G.M. (2006). Cancer stem cells--perspectives on current status and future directions: AACR Workshop on cancer stem cells. Cancer Res. *66*, 9339–9344.
8. Shlush, L.I., Mitchell, A., Heisler, L., Abelson, S., Ng, S.W.K., Trotman-Grant, A., Medeiros, J.J.F., Rao-Bhatia, A., Jaciw-Zurakowsky, I., Marke, R., et al. (2017). Tracing the origins of relapse in acute myeloid leukaemia to stem cells. Nature *547*, 104–108.
9. Dick, J.E. (2008). Stem cell concepts renew cancer research. Blood *112*, 4793–4807.
10. De Grandis, M., Mancini, S.J., and Aurrand-Lions, M. (2018). In quest for leukemia initiating cells in AML. Oncoscience *5*, 9–10.
11. Brandes, R., Lang, F., and Schmidt, R.F. (2010). Physiologie des Menschen (Springer Berlin Heidelberg).
12. Shemin, D., and Rittenberg, D. (1946). The life span of the human red blood cell. J. Biol. Chem. *166*, 627–636.
13. Morrison, S.J., and Kimble, J. (2006). Asymmetric and symmetric stem-cell divisions in development and cancer. Nature *441*, 1068–1074.
14. Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell

biology. Cell *132*, 631–644.

15. Kondo, M. (2010). Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. Immunol. Rev. *238*, 37–46.
16. Seita, J., and Weissman, I.L. (2010). Hematopoietic stem cell: self-renewal versus differentiation. Wiley Interdiscip. Rev. Syst. Biol. Med. *2*, 640–653.
17. AbuSamra, D.B., Aleisa, F.A., Al-Amoodi, A.S., Jalal Ahmed, H.M., Chin, C.J., Abuelela, A.F., Bergam, P., Sougrat, R., and Merzaban, J.S. (2017). Not just a marker: CD34 on human hematopoietic stem/progenitor cells dominates vascular selectin binding along with CD44. Blood Adv. *1*, 2799–2816.
18. Sánchez Alvarado, A., and Yamanaka, S. (2014). Rethinking differentiation: stem cells, regeneration, and plasticity. Cell *157*, 110–119.
19. Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281.
20. Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat. Genet. *48*, 1193–1203.
21. Tikhonova, A.N., Dolgalev, I., Hu, H., Sivaraj, K.K., Hoxha, E., Cuesta-Domínguez, Á., Pinho, S., Akhmetzyanova, I., Gao, J., Witkowski, M., et al. (2019). The bone marrow microenvironment at single-cell resolution. Nature *569*, 222–228.
22. Fröbel, J., Landspersky, T., Percin, G., Schreck, C., Rahmig, S., Ori, A., Nowak, D., Essers, M., Waskow, C., and Oostendorp, R.A.J. (2021). The Hematopoietic Bone Marrow Niche Ecosystem. Front. cell Dev. Biol. *9*, 705410.
23. Morrison, S.J., and Scadden, D.T. (2014). The bone marrow niche for haematopoietic stem cells. Nature *505*, 327–334.
24. Wisniewski, D., Affer, M., Willshire, J., and Clarkson, B. (2011). Further phenotypic characterization of the primitive lineage- CD34+CD38-CD90+CD45RA- hematopoietic stem cell/progenitor cell sub-population isolated from cord blood, mobilized peripheral blood and patients with chronic myelogenous leukemia. Blood Cancer J. *1*, e36.
25. Challen, G.A., Boles, N.C., Chambers, S.M., and Goodell, M.A. (2010). Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1. Cell Stem Cell *6*, 265–278.

26. Yang, L., Bryder, D., Adolfsson, J., Nygren, J., Månsson, R., Sigvardsson, M., and Jacobsen, S.E.W. (2005). Identification of Lin(-)Sca1(+)kit(+)CD34(+)Flt3- short-term hematopoietic stem cells capable of rapidly reconstituting and rescuing myeloablated transplant recipients. Blood *105*, 2717–2723.

27. Haas, S., Trumpp, A., and Milsom, M.D. (2018). Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. Cell Stem Cell *22*, 627–638.

28. Wilson, A., Laurenti, E., Oser, G., van der Wath, R.C., Blanco-Bose, W., Jaworski, M., Offner, S., Dunant, C.F., Eshkind, L., Bockamp, E., et al. (2008). Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. Cell *135*, 1118–1129.

29. Walter, D., Lier, A., Geiselhart, A., Thalheimer, F.B., Huntscha, S., Sobotta, M.C., Moehrle, B., Brocks, D., Bayindir, I., Kaschutnig, P., et al. (2015). Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells. Nature *520*, 549–552.

30. Essers, M.A.G., and Trumpp, A. (2010). Targeting leukemic stem cells by breaking their dormancy. Mol. Oncol. *4*, 443–450.

31. Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391–405.

32. Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M., Gupta, V., Kennedy, J.A., Schimmer, A.D., Schuh, A.C., Yee, K.W., et al. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature *506*, 328–333.

33. Jäger, P., Geyh, S., Twarock, S., Cadeddu, R.-P., Rabes, P., Koch, A., Maus, U., Hesper, T., Zilkens, C., Rautenberg, C., et al. (2021). Acute myeloid leukemia-induced functional inhibition of healthy CD34+ hematopoietic stem and progenitor cells. Stem Cells *39*, 1270–1284.

34. Gruszka, A.M., Valli, D., and Alcalay, M. (2017). Understanding the molecular basis of acute myeloid leukemias: where are we now? Int. J. Hematol. Oncol. *6*, 43–53.

35. Young, A.L., Challen, G.A., Birmann, B.M., and Druley, T.E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun. *7*, 12484.

36. Jaiswal, S., and Ebert, B.L. (2019). Clonal hematopoiesis in human aging and disease. Science *366*.

37. Steensma, D.P., Bejar, R., Jaiswal, S., Lindsley, R.C., Sekeres, M.A., Hasserjian, R.P., and Ebert, B.L. (2015). Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood *126*, 9–16.

38. Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L., Quake, S.R., and Majeti, R. (2012). Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Sci. Transl. Med. *4*, 149ra118.

39. Corces-Zimmerman, M.R., Hong, W.-J., Weissman, I.L., Medeiros, B.C., and Majeti, R. (2014). Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. Proc. Natl. Acad. Sci. U. S. A. *111*, 2548–2553.

40. van Galen, P., Hovestadt, V., Wadsworth, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell *176*, 1265-1281.e24.

41. Thomas, D., and Majeti, R. (2017). Biology and relevance of human acute myeloid leukemia stem cells. Blood *129*, 1577–1585.

42. Dick, J.E., and Bonnet, D. (1997). Human Acute Myeloid Leukaemia is organised as a heirarchy that originates from a primitive haematopoetic cell. Nat. Med. *3*, 730–737.

43. Velten, L., Story, B.A., Hernández-Malmierca, P., Raffel, S., Leonce, D.R., Milbank, J., Paulsen, M., Demir, A., Szu-Tu, C., Frömel, R., et al. (2021). Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. Nat. Commun. *12*, 1366.

44. Löwenberg, B., Downing, J.R., and Burnett, A. (1999). Acute myeloid leukemia. N. Engl. J. Med. *341*, 1051–1062.

45. Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

46. Krause, D.S., Fulzele, K., Catic, A., Sun, C.C., Dombkowski, D., Hurley, M.P., Lezeau, S., Attar, E., Wu, J.Y., Lin, H.Y., et al. (2013). Differential regulation of myeloid leukemias by the bone marrow microenvironment. Nat. Med. *19*, 1513–1517.

47. Riether, C., Schürch, C.M., and Ochsenbein, A.F. (2015). Regulation of hematopoietic and leukemic stem cells by the immune system. Cell Death Differ. *22*, 187–198.

48. Tabe, Y., and Konopleva, M. (2014). Advances in understanding the leukaemia microenvironment. Br. J. Haematol. *164*, 767–778.

49. Kreso, A., and Dick, J.E. (2014). Evolution of the cancer stem cell model. Cell Stem Cell *14*, 275–291.

50. Eppert, K., Takenaka, K., Lechman, E.R., Waldron, L., Nilsson, B., van Galen, P., Metzeler, K.H., Poeppl, A., Ling, V., Beyene, J., et al. (2011). Stem cell gene expression programs influence clinical outcome in human leukemia. Nat. Med. *17*, 1086–1093.

51. Ng, S.W.K., Mitchell, A., Kennedy, J.A., Chen, W.C., McLeod, J., Ibrahimova, N., Arruda, A., Popescu, A., Gupta, V., Schimmer, A.D., et al. (2016). A 17-gene stemness score for rapid determination of risk in acute leukaemia. Nature *540*, 433–437.

52. Jin, L., Lee, E.M., Ramshaw, H.S., Busfield, S.J., Peoppl, A.G., Wilkinson, L., Guthridge, M.A., Thomas, D., Barry, E.F., Boyd, A., et al. (2009). Monoclonal antibody-mediated targeting of CD123, IL-3 receptor alpha chain, eliminates human acute myeloid leukemic stem cells. Cell Stem Cell *5*, 31–42.

53. Majeti, R., Chao, M.P., Alizadeh, A.A., Pang, W.W., Jaiswal, S., Gibbs, K.D.J., van Rooijen, N., and Weissman, I.L. (2009). CD47 is an adverse prognostic factor and therapeutic antibody target on human acute myeloid leukemia stem cells. Cell *138*, 286–299.

54. Kikushige, Y., Shima, T., Takayanagi, S., Urata, S., Miyamoto, T., Iwasaki, H., Takenaka, K., Teshima, T., Tanaka, T., Inagaki, Y., et al. (2010). TIM-3 is a promising target to selectively kill acute myeloid leukemia stem cells. Cell Stem Cell *7*, 708–717.

55. Gerber, J.M., Smith, B.D., Ngwang, B., Zhang, H., Vala, M.S., Morsberger, L., Galkin, S., Collector, M.I., Perkins, B., Levis, M.J., et al. (2012). A clinically relevant population of leukemic CD34(+)CD38(-) cells in acute myeloid leukemia. Blood *119*, 3571–3577.

56. Saito, Y., Kitamura, H., Hijikata, A., Tomizawa-Murasawa, M., Tanaka, S., Takagi, S., Uchida, N., Suzuki, N., Sone, A., Najima, Y., et al. (2010). Identification of therapeutic targets for quiescent, chemotherapy-resistant human leukemia stem cells. Sci. Transl. Med. *2*, 17ra9.

57. Mazzarella, L., Riva, L., Luzi, L., Ronchini, C., and Pelicci, P.G. (2014). The Genomic and Epigenomic Landscapes of AML. Semin. Hematol. *51*, 259–272.

58. Assi, S.A., Imperato, M.R., Coleman, D.J.L., Pickin, A., Potluri, S., Ptasinska, A., Chin, P.S., Blair, H., Cauchy, P., James, S.R., et al. (2019). Subtype-specific regulatory network rewiring in acute myeloid leukemia. Nat. Genet. *51*, 151–162.

59. Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J.J., Laird, P.W., Baty, J.D., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

60. Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–

510.

61. Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

62. Spencer, D.H., Russler-Germain, D.A., Ketkar, S., Helton, N.M., Lamprecht, T.L., Fulton, R.S., Fronick, C.C., O'Laughlin, M., Heath, S.E., Shinawi, M., et al. (2017). CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. Cell *168*, 801-816.e13.

63. Sandoval, J.E., Huang, Y.-H., Muise, A., Goodell, M.A., and Reich, N.O. (2019). Mutations in the DNMT3A DNA methyltransferase in acute myeloid leukemia patients cause both loss and gain of function and differential regulation by protein partners. J. Biol. Chem. *294*, 4898–4910.

64. Shih, A.H., Abdel-Wahab, O., Patel, J.P., and Levine, R.L. (2012). The role of mutations in epigenetic regulators in myeloid malignancies. Nat. Rev. Cancer *12*, 599–612.

65. Brunetti, L., Gundry, M.C., and Goodell, M.A. (2017). DNMT3A in Leukemia. Cold Spring Harb. Perspect. Med. *7*.

66. Jeong, M., Park, H.J., Celik, H., Ostrander, E.L., Reyes, J.M., Guzman, A., Rodriguez, B., Lei, Y., Lee, Y., Ding, L., et al. (2018). Loss of Dnmt3a Immortalizes Hematopoietic Stem Cells In Vivo. Cell Rep. *23*, 1–10.

67. Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., Kandoth, C., Payton, J.E., Baty, J., Welch, J., et al. (2010). DNMT3A mutations in acute myeloid leukemia. N. Engl. J. Med. *363*, 2424–2433.

68. Gale, R.E., Lamb, K., Allen, C., El-Sharkawi, D., Stowe, C., Jenkinson, S., Tinsley, S., Dickson, G., Burnett, A.K., Hills, R.K., et al. (2015). Simpson's Paradox and the Impact of Different DNMT3A Mutations on Outcome in Younger Adults With Acute Myeloid Leukemia. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *33*, 2072–2083.

69. Gaidzik, V.I., Schlenk, R.F., Paschka, P., Stölzle, A., Späth, D., Kuendgen, A., von Lilienfeld-Toal, M., Brugger, W., Derigs, H.G., Kremers, S., et al. (2013). Clinical impact of DNMT3A mutations in younger adult patients with acute myeloid leukemia: results of the AML Study Group (AMLSG). Blood *121*, 4769–4777.

70. Falini, B., Mecucci, C., Tiacci, E., Alcalay, M., Rosati, R., Pasqualucci, L., La Starza, R., Diverio, D., Colombo, E., Santucci, A., et al. (2005). Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. N. Engl. J. Med. *352*, 254–266.

71. Wang, W., Budhu, A., Forgues, M., and Wang, X.W. (2005). Temporal and spatial

control of nucleophosmin by the Ran-Crm1 complex in centrosome duplication. Nat. Cell Biol. *7*, 823–830.

72. Heath, E.M., Chan, S.M., Murphy, T., Shlush, L.I., and Schimmer, A.D. (2017). Biological and clinical consequences of NPM1 mutations in AML. Nat. Publ. Gr., 798–807.
73. Colombo, E., Martinelli, P., Zamponi, R., Shing, D.C., Bonetti, P., Luzi, L., Volorio, S., Bernard, L., Pruneri, G., Alcalay, M., et al. (2006). Delocalization and destabilization of the Arf tumor suppressor by the leukemia-associated NPM mutant. Cancer Res. *66*, 3044–3050.
74. Grisendi, S., Bernardi, R., Rossi, M., Cheng, K., Khandker, L., Manova, K., and Pandolfi, P.P. (2005). Role of nucleophosmin in embryonic development and tumorigenesis. Nature *437*, 147–153.
75. Döhner, K., Schlenk, R.F., Habdank, M., Scholl, C., Rücker, F.G., Corbacioglu, A., Bullinger, L., Fröhling, S., and Döhner, H. (2005). Mutant nucleophosmin (NPM1) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: interaction with other gene mutations. Blood *106*, 3740–3746.
76. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. N. Engl. J. Med. *374*, 2209–2221.
77. Daneshbod, Y., Kohan, L., Taghadosi, V., Weinberg, O.K., and Arber, D.A. (2019). Prognostic Significance of Complex Karyotypes in Acute Myeloid Leukemia. Curr. Treat. Options Oncol. *20*, 15.
78. Ferrara, F., and Schiffer, C.A. (2013). Acute myeloid leukaemia in adults. Lancet (London, England) *381*, 484–95.
79. Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F.R., Büchner, T., Dombret, H., Ebert, B.L., Fenaux, P., Larson, R.A., et al. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood *129*, 424–447.
80. Goldstone, A.H., Burnett, A.K., Wheatley, K., Smith, A.G., Hutchinson, R.M., and Clark, R.E. (2001). Attempts to improve treatment outcomes in acute myeloid leukemia (AML) in older patients: the results of the United Kingdom Medical Research Council AML11 trial. Blood *98*, 1302–1311.
81. Biswal, S., and Godnaik, C. (2013). Incidence and management of infections in patients with acute leukemia following chemotherapy in general wards. Ecancermedicalscience *7*, 310.

82. DiNardo, C.D., Pratz, K., Pullarkat, V., Jonas, B.A., Arellano, M., Becker, P.S., Frankfurt, O., Konopleva, M., Wei, A.H., Kantarjian, H.M., et al. (2019). Venetoclax combined with decitabine or azacitidine in treatment-naive, elderly patients with acute myeloid leukemia. Blood *133*, 7–17.

83. Guerra, V.A., DiNardo, C., and Konopleva, M. (2019). Venetoclax-based therapies for acute myeloid leukemia. Best Pract. Res. Clin. Haematol. *32*, 145–153.

84. Thol, F., and Ganser, A. (2020). Treatment of Relapsed Acute Myeloid Leukemia. Curr. Treat. Options Oncol. *21*, 66.

85. Jones, C.L., Stevens, B.M., D'Alessandro, A., Reisz, J.A., Culp-Hill, R., Nemkov, T., Pei, S., Khan, N., Adane, B., Ye, H., et al. (2018). Inhibition of Amino Acid Metabolism Selectively Targets Human Leukemia Stem Cells. Cancer Cell *34*, 724-740.e4.

86. Pollyea, D.A., Stevens, B.M., Jones, C.L., Winters, A., Pei, S., Minhajuddin, M., D'Alessandro, A., Culp-Hill, R., Riemondy, K.A., Gillen, A.E., et al. (2018). Venetoclax with azacitidine disrupts energy metabolism and targets leukemia stem cells in patients with acute myeloid leukemia. Nat. Med. *24*, 1859–1866.

87. Martiáñez Canales, T., de Leeuw, D.C., Vermue, E., Ossenkoppele, G.J., and Smit, L. (2017). Specific Depletion of Leukemic Stem Cells: Can MicroRNAs Make the Difference? Cancers (Basel). *9*.

88. Essers, M.A.G., Offner, S., Blanco-Bose, W.E., Waibler, Z., Kalinke, U., Duchosal, M.A., and Trumpp, A. (2009). IFNalpha activates dormant haematopoietic stem cells in vivo. Nature *458*, 904–908.

89. Velasco-Hernandez, T., Soneji, S., Hidalgo, I., Erlandsson, E., Cammenga, J., and Bryder, D. (2019). Hif-1α Deletion May Lead to Adverse Treatment Effect in a Mouse Model of MLL-AF9-Driven AML. Stem cell reports *12*, 112–121.

90. Tyner, J.W., Tognon, C.E., Bottomly, D., Wilmot, B., Kurtz, S.E., Savage, S.L., Long, N., Schultz, A.R., Traer, E., Abel, M., et al. (2018). Functional genomic landscape of acute myeloid leukaemia. Nature *562*, 526–531.

91. Lapidot, T., Sirard, C., Vormoor, J., Murdoch, B., Hoang, T., Caceres-Cortes, J., Minden, M., Paterson, B., Caligiuri, M.A., and Dick, J.E. (1994). A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. Nature *367*, 645–648.

92. Dick, J.E., and Lapidot, T. (2005). Biology of normal and acute myeloid leukemia stem cells. Int. J. Hematol. *82*, 389–396.

93. Taussig, D.C., Vargaftig, J., Miraki-Moud, F., Griessinger, E., Sharrock, K., Luke, T., Lillington, D., Oakervee, H., Cavenagh, J., Agrawal, S.G., et al. (2010). Leukemia-

initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the CD34- fraction. Blood *115*, 1976–1984.

94. Pabst, C., Bergeron, A., Lavallée, V.P., Yeh, J., Gendron, P., Norddahl, G.L., Krosl, J., Boivin, I., Deneault, E., Simard, J., et al. (2016). GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. Blood *127*, 2018–2027.
95. Singh, A.K., and Lin, H.-H. (2021). The role of GPR56/ADGRG1 in health and disease. Biomed. J. *44*, 534–547.
96. Rao, T.N., Marks-Bluth, J., Sullivan, J., Gupta, M.K., Chandrakanthan, V., Fitch, S.R., Ottersbach, K., Jang, Y.C., Piao, X., Kulkarni, R.N., et al. (2015). High-level Gpr56 expression is dispensable for the maintenance and function of hematopoietic stem and progenitor cells in mice. Stem Cell Res. *14*, 307–322.
97. Peng, Y.-M., van de Garde, M.D.B., Cheng, K.-F., Baars, P.A., Remmerswaal, E.B.M., van Lier, R.A.W., Mackay, C.R., Lin, H.-H., and Hamann, J. (2011). Specific expression of GPR56 by human cytotoxic lymphocytes. J. Leukoc. Biol. *90*, 735–740.
98. Paczulla, A.M., Rothfelder, K., Raffel, S., Konantz, M., Steinbacher, J., Wang, H., Tandler, C., Mbarga, M., Schaefer, T., Falcone, M., et al. (2019). Absence of NKG2D ligands defines leukaemia stem cells and mediates their immune evasion. Nature *572*, 254–259.
99. Raulet, D.H., Gasser, S., Gowen, B.G., Deng, W., and Jung, H. (2013). Regulation of Ligands for the NKG2D Activating Receptor. Annu. Rev. Immunol. *31*, 413–441.
100. Zingoni, A., Molfetta, R., Fionda, C., Soriani, A., Paolini, R., Cippitelli, M., Cerboni, C., and Santoni, A. (2018). NKG2D and Its Ligands: “One for All, All for One”. Front. Immunol. *9*, 476.
101. Martelli, M.P., Pettirossi, V., Thiede, C., Bonifacio, E., Mezzasoma, F., Cecchini, D., Pacini, R., Tabarrini, A., Ciurnelli, R., Gionfriddo, I., et al. (2010). CD34+ cells from AML with mutated NPM1 harbor cytoplasmic mutated nucleophosmin and generate leukemia in immunocompromised mice. Blood *116*, 3907–3922.
102. Alberts, B., Johnson, A., Wilson, J., Lewis, J., Hunt, T., Roberts, K., Raff, M., and Walter, P. (2008). Molecular Biology of the Cell (Garland Science).
103. You, J.S., and Jones, P.A. (2012). Cancer Genetics and Epigenetics: Two Sides of the Same Coin? Cancer Cell *22*, 9–20.
104. Yang, L., Rau, R., and Goodell, M.A. (2015). DNMT3A in haematological malignancies. Nat. Rev. Cancer *15*, 152–165.
105. Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription.

Genes Dev. *25*, 1010–1022.

106. Cross, S.H., and Bird, A.P. (1995). CpG islands and genes. Curr. Opin. Genet. Dev. *5*, 309–314.
107. Kim, M., and Costello, J. (2017). DNA methylation: An epigenetic mark of cellular memory. Exp. Mol. Med. *49*.
108. Ji, H., Ehrlich, L.I.R., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M.J., Irizarry, R.A., Kim, K., et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature *467*, 338–342.
109. Álvarez-Errico, D., Vento-Tormo, R., Sieweke, M., and Ballestar, E. (2015). Epigenetic control of myeloid cell differentiation, identity and function. Nat. Rev. Immunol. *15*, 7–17.
110. Klutstein, M., Nejman, D., Greenfield, R., and Cedar, H. (2016). DNA methylation in cancer and aging. Cancer Res. *76*, 3446–3450.
111. Schoofs, T., and Müller-Tidow, C. (2011). DNA methylation as a pathogenic event and as a therapeutic target in AML. Cancer Treat. Rev. *37*, S13–S18.
112. Gallego-Paez, L.M., Bordone, M.C., Leote, A.C., Saraiva-Agostinho, N., Ascensão-Ferreira, M., and Barbosa-Morais, N.L. (2017). Alternative splicing: the pledge, the turn, and the prestige : The key role of alternative splicing in human biological systems. Hum. Genet. *136*, 1015–1042.
113. Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. *18*, 437–451.
114. Wang, Y., Liu, J., Huang, B.O., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M., et al. (2015). Mechanism of alternative splicing and its regulation. Biomed. reports *3*, 152–158.
115. Coelho, M.B., and Smith, C.W.J. (2014). Regulation of alternative pre-mRNA splicing. Methods Mol. Biol. *1126*, 55–82.
116. Lev Maor, G., Yearim, A., and Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. Trends Genet. *31*, 274–280.
117. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74–79.
118. Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U. S. A. *111*, E5593–E5601.

119. Ehx, G., Larouche, J.D., Durette, C., Laverdure, J.P., Hesnard, L., Vincent, K., Hardy, M.P., Thériault, C., Rulleau, C., Lanoix, J., et al. (2021). Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. Immunity *54*, 737-752.e10.

120. Jin, P., Tan, Y., Zhang, W., Li, J., and Wang, K. (2020). Prognostic alternative mRNA splicing signatures and associated splicing factors in acute myeloid leukemia. Neoplasia (United States) *22*, 447–457.

121. Kim, E., Ilagan, J.O., Liang, Y., Daubner, G.M., Lee, S.C.-W., Ramakrishnan, A., Li, Y., Chung, Y.R., Micol, J.-B., Murphy, M.E., et al. (2015). SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. Cancer Cell *27*, 617–630.

122. Hershberger, C.E., Moyer, D.C., Adema, V., Kerr, C.M., Walter, W., Hutter, S., Meggendorfer, M., Baer, C., Kern, W., Nadarajah, N., et al. (2021). Complex landscape of alternative splicing in myeloid neoplasms. Leukemia, 1108–1120.

123. Danckwardt, S., Neu-Yilik, G., Thermann, R., Frede, U., Hentze, M.W., and Kulozik, A.E. (2002). Abnormally spliced beta-globin mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay. Blood *99*, 1811–1816.

124. Banaszak, L.G., Giudice, V., Zhao, X., Wu, Z., Gao, S., Hosokawa, K., Keyvanfar, K., Townsley, D.M., Gutierrez-Rodrigues, F., Fernandez Ibanez, M. del P., et al. (2018). Abnormal RNA splicing and genomic instability after induction of DNMT3A mutations by CRISPR/Cas9 gene editing. Blood Cells, Mol. Dis. *69*, 10–22.

125. de Necochea-Campion, R., Shouse, G.P., Zhou, Q., Mirshahidi, S., and Chen, C.-S. (2016). Aberrant splicing and drug resistance in AML. J. Hematol. Oncol. *9*, 85.

126. Rivera, O.D., Mallory, M.J., Quesnel-Vallières, M., Chatrikhi, R., Schultz, D.C., Carroll, M., Barash, Y., Cherry, S., and Lynch, K.W. (2021). Alternative splicing redefines landscape of commonly mutated genes in acute myeloid leukemia. Proc. Natl. Acad. Sci. *118*, e2014967118.

127. Zhang, N., Zhang, P., Chen, Y., Lou, S., Zeng, H., and Deng, J. (2020). Clusterization in acute myeloid leukemia based on prognostic alternative splicing signature to reveal the clinical characteristics in the bone marrow microenvironment. Cell Biosci. *10*, 1–14.

128. Zhang, B., Yang, L., Wang, X., and Fu, D. (2021). Identification of Survival-related Alternative Splicing Signatures in Acute Myeloid Leukemia. Biosci. Rep.

129. Anande, G., Deshpande, N.P., Mareschal, S., Batcha, A.M.N., Hampton, H.R., Herold, T.,

Lehmann, S., Wilkins, M.R., Wong, J.W.H., Unnikrishnan, A., et al. (2020). RNA Splicing Alterations Induce a Cellular Stress Response Associated with Poor Prognosis in Acute Myeloid Leukemia. Clin. Cancer Res. *26*, 3597–3607.

130. Bowman, T. V. (2020). Improving AML Classification Using Splicing Signatures. Clin. Cancer Res. *26*, 3503–3504.
131. Merino, G.A., Conesa, A., and Fernández, E.A. (2019). A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. Brief. Bioinform. *20*, 471–481.
132. Mehmood, A., Laiho, A., Venälaïnen, M.S., McGlinchey, A.J., Wang, N., and Elo, L.L. (2020). Systematic evaluation of differential splicing tools for RNA-seq studies. Brief. Bioinform. *21*, 2052–2065.
133. Gallego-Paez, L.M., and Mauer, J. (2022). DJExpress: An Integrated Application for Differential Splicing Analysis and Visualization . Front. Bioinforma. *2*.
134. Worley, B., and Powers, R. (2012). Multivariate Analysis in Metabolomics. Curr. Metabolomics *1*, 92–107.
135. Vinayavekhin, N., Homan, E.A., and Saghatelian, A. (2010). Exploring Disease through Metabolomics. ACS Chem. Biol. *5*, 91–103.
136. Wilcoxen, K.M., Uehara, T., Myint, K.T., Sato, Y., and Oda, Y. (2010). Practical metabolomics in drug discovery. Expert Opin. Drug Discov. *5*, 249–263.
137. Wickenhagen, W. V., Salomons, G.S., Gibson, K.M., Jakobs, C., and Struys, E.A. (2009). Measurement of D-2-hydroxyglutarate dehydrogenase activity in cell homogenates derived from D-2-hydroxyglutaric aciduria patients. J. Inherit. Metab. Dis. *32*, 264–268.
138. Wilson, M.P., Footitt, E.J., Papandreou, A., Uudelepp, M.L., Pressler, R., Stevenson, D.C., Gabriel, C., McSweeney, M., Baggot, M., Burke, D., et al. (2017). An LC-MS/MS-Based Method for the Quantification of Pyridox(am)ine 5'-Phosphate Oxidase Activity in Dried Blood Spots from Patients with Epilepsy. Anal. Chem. *89*, 8892–8900.
139. Peters, T.M.A., Engelke, U.F.H., de Boer, S., van der Heeft, E., Pritsch, C., Kulkarni, P., Wevers, R.A., Willemsen, M.A.A.P., Verbeek, M.M., and Coene, K.L.M. (2020). Confirmation of neurometabolic diagnoses using age-dependent cerebrospinal fluid metabolomic profiles. J. Inherit. Metab. Dis. *43*, 1112–1120.
140. Sitole, L.J., Williams, A.A., and Meyer, D. (2013). Metabonomic analysis of HIV-infected biofluids. Mol. Biosyst. *9*, 18–28.
141. Chandler, J.D., Hu, X., Ko, E.J., Park, S., Lee, Y.T., Orr, M., Fernandes, J., Uppal, K., Kang, S.M., Jones, D.P., et al. (2016). Metabolic pathways of lung inflammation revealed by

high-resolution metabolomics (HRM) of H1N1 influenza virus infection in mice. Am. J. Physiol. - Regul. Integr. Comp. Physiol. *311*, R906–R916.

142. Aurpibul, L., Namwongprom, S., Sudjaritruk, T., and Ounjaijean, S. (2020). Metabolic syndrome, biochemical markers, and body composition in youth living with perinatal HIV infection on antiretroviral treatment. PLoS One *15*, 1–13.
143. Nagy-Szakal, D., Williams, B.L., Mishra, N., Che, X., Lee, B., Bateman, L., Klimas, N.G., Komaroff, A.L., Levine, S., Montoya, J.G., et al. (2017). Fecal metagenomic profiles in subgroups of patients with myalgic encephalomyelitis/chronic fatigue syndrome. Microbiome *5*, 44.
144. Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A.L. (2019). Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature *570*, 462–467.
145. Mimmi, M.C., Picotti, P., Corazza, A., Betto, E., Pucillo, C.E., Cesaratto, L., Cedolini, C., Londero, V., Zuiani, C., Bazzocchi, M., et al. (2011). High-performance metabolic marker assessment in breast cancer tissue by mass spectrometry. Clin. Chem. Lab. Med. *49*, 317–324.
146. Banerjee, S., Zare, R.N., Tibshirani, R.J., Kunder, C.A., Nolley, R., Fan, R., Brooks, J.D., and Sonn, G.A. (2017). Diagnosis of prostate cancer by desorption electrospray ionization mass spectrometric imaging of small metabolites and lipids. Proc. Natl. Acad. Sci. U. S. A. *114*, 3334–3339.
147. Zukunft, S., Prehn, C., Röhring, C., Möller, G., Hrabě de Angelis, M., Adamski, J., and Tokarz, J. (2018). High-throughput extraction and quantification method for targeted metabolomics in murine tissues. Metabolomics *14*, 1–12.
148. Ivanisevic, J., Zhu, Z.J., Plate, L., Tautenhahn, R., Chen, S., O'Brien, P.J., Johnson, C.H., Marletta, M.A., Patti, G.J., and Siuzdak, G. (2013). Toward 'Omic scale metabolite profiling: A dual separation-mass spectrometry approach for coverage of lipid and central carbon metabolism. Anal. Chem. *85*, 6876–6884.
149. Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D., and McLean, J.A. (2016). Untargeted Metabolomics Strategies-Challenges and Emerging Directions. J. Am. Soc. Mass Spectrom. *27*, 1897–1905.
150. Burla, B., Arita, M., Arita, M., Bendt, A.K., Cazenave-Gassiot, A., Dennis, E.A., Ekroos, K., Han, X., Ikeda, K., Liebisch, G., et al. (2018). MS-based lipidomics of human blood plasma: A community-initiated position paper to develop accepted guidelines. J. Lipid Res. *59*, 2001–2017.

151. Blennow, K., and Zetterberg, H. (2018). Biomarkers for Alzheimer's disease: current status and prospects for the future. J. Intern. Med. *284*, 643–663.

152. Zapf, A., Gwinner, W., Karch, A., Metzger, J., Haller, H., and Koch, A. (2015). Non-invasive diagnosis of acute rejection in renal transplant patients using mass spectrometry of urine samples - A multicentre phase 3 diagnostic accuracy study. BMC Nephrol. *16*, 1–7.

153. Harpole, M., Davis, J., and Espina, V. (2016). Expert Review of Proteomics Current state of the art for enhancing urine biomarker discovery Current state of the art for enhancing urine biomarker discovery. Expert Rev Proteomics *9450*, 609–626.

154. Gowda, G.A.N. (2018). Profiling Redox and Energy Coenzymes in Whole Blood, Tissue and Cells Using NMR Spectroscopy. Metabolites *8*.

155. Yin, P., Lehmann, R., and Xu, G. (2015). Effects of pre-analytical processes on blood samples used in metabolomics studies. Anal. Bioanal. Chem. *407*, 4879–4892.

156. Khalid, A., Siddiqui, A.J., Huang, J.-H., Shamsi, T., and Musharraf, S.G. (2018). Alteration of Serum Free Fatty Acids are Indicators for Progression of Pre-leukaemia Diseases to Leukaemia. Sci. Rep. *8*, 14883.

157. Agathocleous, M., Meacham, C.E., Burgess, R.J., Piskounova, E., Zhao, Z., Crane, G.M., Cowin, B.L., Bruner, E., Murphy, M.M., Chen, W., et al. (2017). Ascorbate regulates haematopoietic stem cell function and leukaemogenesis. Nature *549*, 476–481.

158. Darici, S., Alkhaldi, H., Horne, G., Jørgensen, H.G., Marmiroli, S., and Huang, X. (2020). Targeting pi3k/akt/mtor in aml: Rationale and clinical evidence. J. Clin. Med. *9*, 1–40.

159. Donato, E., and Trumpp, A. (2022). Targeting the Leukemic stem cell protein machinery by inhibition of mitochondrial pyrimidine synthesis. EMBO Mol. Med. *14*, e16171.

160. Simsek, T., Kocabas, F., Zheng, J., Deberardinis, R.J., Mahmoud, A.I., Olson, E.N., Schneider, J.W., Zhang, C.C., and Sadek, H.A. (2010). The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche. Cell Stem Cell *7*, 380–390.

161. Erdem, A., Marin, S., Pereira-Martins, D.A., Cortés, R., Cunningham, A., Pruis, M.G., de Boer, B., van den Heuvel, F.A.J., Geugien, M., Wierenga, A.T.J., et al. (2022). The Glycolytic Gatekeeper PDK1 defines different metabolic states between genetically distinct subtypes of human acute myeloid leukemia. Nat. Commun. *13*.

162. Lagadinou, E.D., Sach, A., Callahan, K., Rossi, R.M., Neering, S.J., Minhajuddin, M., Ashton, J.M., Pei, S., Grose, V., O'Dwyer, K.M., et al. (2013). BCL-2 inhibition targets oxidative phosphorylation and selectively eradicates quiescent human leukemia stem

cells. Cell Stem Cell *12*, 329–341.

163. de Beauchamp, L., Himonas, E., and Helgason, G.V. (2022). Mitochondrial metabolism as a potential therapeutic target in myeloid leukaemia. Leukemia *36*, 1–12.
164. Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discov. *12*, 31–46.
165. Fasouli, E.S., and Katsantoni, E. (2021). JAK-STAT in Early Hematopoiesis and Leukemia . Front. Cell Dev. Biol. *9*.
166. Massagué, J. (2008). TGFbeta in Cancer. Cell *134*, 215–230.
167. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Cell *144*, 646–674.
168. Coller, H.A. (2011). Cell biology. The essence of quiescence. Science *334*, 1074–1075.
169. Welch, J.S. (2018). Patterns of mutations in TP53 mutated AML. Best Pract. Res. Clin. Haematol. *31*, 379–383.
170. Kreitz, J., Schönfeld, C., Seibert, M., Stolp, V., Alshamleh, I., Oellerich, T., Steffen, B., Schwalbe, H., Schnütgen, F., Kurrle, N., et al. (2019). Metabolic Plasticity of Acute Myeloid Leukemia. Cells *8*, 805.
171. Fadaka, A., Ajiboye, B., Ojo, O., Adewale, O., Olayide, I., and Emuowhochere, R. (2017). Biology of glucose metabolization in cancer cells. J. Oncol. Sci. *3*, 45–51.
172. Vander Heiden, M.G., Cantley, L.C., and Thompson, C.B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science *324*, 1029–1033.
173. Gandini, S., Massi, D., and Mandalà, M. (2016). PD-L1 expression in cancer patients receiving anti PD-1/PD-L1 antibodies: A systematic review and meta-analysis. Crit. Rev. Oncol. Hematol. *100*, 88–98.
174. Syn, N.L., Teng, M.W.L., Mok, T.S.K., and Soo, R.A. (2017). De-novo and acquired resistance to immune checkpoint targeting. Lancet. Oncol. *18*, e731–e741.
175. Stahl, M., and Goldberg, A.D. (2019). Immune Checkpoint Inhibitors in Acute Myeloid Leukemia: Novel Combinations and Therapeutic Targets. Curr. Oncol. Rep. *21*, 37.
176. Hernández-Malmierca, P., Vonficht, D., Schnell, A., Uckelmann, H.J., Bollhagen, A., Mahmoud, M.A.A., Landua, S.L., van der Salm, E., Trautmann, C.L., Raffel, S., et al. (2022). Antigen presentation safeguards the integrity of the hematopoietic stem cell pool. Cell Stem Cell *29*, 760-775.e10.
177. Toffalori, C., Zito, L., Gambacorta, V., Riba, M., Oliveira, G., Bucci, G., Barcella, M., Spinelli, O., Greco, R., Crucitti, L., et al. (2019). Immune signature drives leukemia escape and relapse after hematopoietic cell transplantation. Nat. Med. *25*, 603–611.

178. Gaipa, G., Coustan-Smith, E., Todisco, E., Maglia, O., Biondi, A., and Campana, D. (2002). Characterization of CD34+, CD13+, CD33- cells, a rare subset of immature human hematopoietic cells. Haematologica *87*, 347–356.

179. Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M., and Campbell, J.D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. *21*, 1–15.

180. Gaidatzis, D., Burger, L., Florescu, M., and Stadler, M.B. (2015). Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat. Biotechnol. *33*, 722–729.

181. Novak, D., Hüser, L., Elton, J.J., Umansky, V., Altevogt, P., and Utikal, J. (2020). SOX2 in development and cancer biology. Semin. Cancer Biol. *67*, 74–82.

182. Jaatinen, T., Hemmoranta, H., Hautaniemi, S., Niemi, J., Nicorici, D., Laine, J., Yli-Harja, O., and Partanen, J. (2006). Global gene expression profile of human cord blood-derived CD133+ cells. Stem Cells *24*, 631–641.

183. Zhang, J., Lee, E.Y., Liu, Y., Berman, S.D., Lodish, H.F., and Lees, J.A. (2010). pRB and E2F4 play distinct cell-intrinsic roles in fetal erythropoiesis. Cell Cycle *9*, 371–376.

184. Hu, X., and Zuckerman, K.S. (2001). Transforming growth factor: signal transduction pathways, cell cycle mediation, and effects on hematopoiesis. J. Hematother. Stem Cell Res. *10*, 67–74.

185. Stevens, J.D., Roalson, E.H., and Skinner, M.K. (2008). Phylogenetic and expression analysis of the basic helix-loop-helix transcription factor gene family: genomic approach to cellular differentiation. Differentiation. *76*, 1006–1022.

186. Vaure, C., and Liu, Y. (2014). A comparative review of toll-like receptor 4 expression and functionality in different animal species. Front. Immunol. *5*, 316.

187. Duy, C., Yu, J.J., Nahar, R., Swaminathan, S., Kweon, S.-M., Polo, J.M., Valls, E., Klemm, L., Shojaee, S., Cerchietti, L., et al. (2010). BCL6 is critical for the development of a diverse primary B cell repertoire. J. Exp. Med. *207*, 1209–1221.

188. Ayoub, E., Wilson, M.P., McGrath, K.E., Li, A.J., Frisch, B.J., Palis, J., Calvi, L.M., Zhang, Y., and Perkins, A.S. (2018). EVI1 overexpression reprograms hematopoiesis via upregulation of Spi1 transcription. Nat. Commun. *9*, 4239.

189. DiNardo, C.D., Garcia-Manero, G., Pierce, S., Nazha, A., Bueso-Ramos, C., Jabbour, E., Ravandi, F., Cortes, J., and Kantarjian, H. (2016). Interactions and relevance of blast percentage and treatment strategy among younger and older patients with acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS). Am. J. Hematol. *91*,

227–232.

190. Naka, K., and Hirao, A. (2017). Regulation of hematopoiesis and hematological disease by TGF-β family signaling molecules. Cold Spring Harb. Perspect. Biol. *9*, 25.

191. Bolouri, H., Farrar, J.E., Triche, T., Ries, R.E., Lim, E.L., Alonzo, T.A., Ma, Y., Moore, R., Mungall, A.J., Marra, M.A., et al. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat. Med. *24*, 103–112.

192. Jayavelu, A.K., Wolf, S., Buettner, F., Alexe, G., Häupl, B., Comoglio, F., Schneider, C., Doebele, C., Fuhrmann, D.C., Wagner, S., et al. (2022). The proteogenomic subtypes of acute myeloid leukemia. Cancer Cell *40*, 301-317.e12.

193. Metzeler, K.H., Hummel, M., Bloomfield, C.D., Spiekermann, K., Braess, J., Sauerland, M.C., Heinecke, A., Radmacher, M., Marcucci, G., Whitman, S.P., et al. (2008). An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. Blood *112*, 4193–4201.

194. Villard, J., Reith, W., Barras, E., Gos, A., Morris, M.A., Antonarakis, S.E., Van Den Elsen, P.J., and Mach, B. (1997). Analysis of mutations and chromosomal localisation of the gene encoding RFX5, a novel transcription factor affected in major histocompatibility complex class II deficiency. Hum. Mutat. *10*, 430–435.

195. Wang, K., Sanchez-Martin, M., Wang, X., Knapp, K.M., Koche, R., Vu, L., Nahas, M.K., He, J., Hadler, M., Stein, E.M., et al. (2017). Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. Leukemia *31*, 151–158.

196. Jung, N., Dai, B., Gentles, A.J., Majeti, R., and Feinberg, A.P. (2015). An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. Nat. Commun. *6*.

197. Hodges, E., Molaro, A., Dos Santos, C.O., Thekkat, P., Song, Q., Uren, P.J., Park, J., Butler, J., Rafii, S., McCombie, W.R., et al. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol. Cell *44*, 17–28.

198. Goardon, N., Marchi, E., Atzberger, A., Quek, L., Schuh, A., Soneji, S., Woll, P., Mead, A., Alford, K.A., Rout, R., et al. (2011). Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. Cancer Cell *19*, 138–152.

199. Bluthgen, M.V., Basté, N., and Recondo, G. (2020). Immunotherapy combinations for the treatment of patients with solid tumors. Future Oncol. *16*, 1715–1736.

200. Isidori, A., Cerchione, C., Daver, N., DiNardo, C., Garcia-Manero, G., Konopleva, M.,

Jabbour, E., Ravandi, F., Kadia, T., Burguera, A. de la F., et al. (2021). Immunotherapy in Acute Myeloid Leukemia: Where We Stand. Front. Oncol. *11*, 656218.

201. Deng, M., Gui, X., Kim, J., Xie, L., Chen, W., Li, Z., He, L., Chen, Y., Chen, H., Luo, W., et al. (2018). LILRB4 signalling in leukaemia cells mediates T cell suppression and tumour infiltration. Nature *562*, 605–609.
202. Blunt, M.D., and Khakoo, S.I. (2020). Activating killer cell immunoglobulin-like receptors: Detection, function and therapeutic use. Int. J. Immunogenet. *47*, 1–12.
203. Johansson, H., and Simonsson, S. (2010). Core transcription factors, Oct4, Sox2 and Nanog, individually form complexes with nucleophosmin (Npm1) to control embryonic stem (ES) cell fate determination. Aging (Albany. NY). *2*, 815–822.
204. Tosic, N., Petrovic, I., Grujicic, N.K., Davidovic, S., Virijevic, M., Vukovic, N.S., Pavlovic, S., and Stevanovic, M. (2018). Prognostic significance of SOX2, SOX3, SOX11, SOX14 and SOX18 gene expression in adult de novo acute myeloid leukemia. Leuk. Res. *67*, 32–38.
205. Bousoik, E., and Montazeri Aliabadi, H. (2018). "Do We Know Jack" About JAK? A Closer Look at JAK/STAT Signaling Pathway. Front. Oncol. *8*, 1–20.
206. Thomas, S.J., Snowden, J.A., Zeidler, M.P., and Danson, S.J. (2015). The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours. Br. J. Cancer *113*, 365–371.
207. SanMiguel, J.M., Eudy, E., Loberg, M.A., Miles, L.A., Stearns, T., Mistry, J.J., Rauh, M.J., Levine, R.L., and Trowbridge, J.J. (2022). Cell origin–dependent cooperativity of mutant Dnmt3a and Npm1 in clonal hematopoiesis and myeloid malignancy . Blood Adv. *6*, 3666–3677.
208. Antoniani, C., Romano, O., and Miccio, A. (2017). Concise Review: Epigenetic Regulation of Hematopoiesis: Biological Insights and Therapeutic Applications. Stem Cells Transl. Med. *6*, 2106–2114.
209. Quek, L., Otto, G.W., Garnett, C., Lhermitte, L., Karamitros, D., Stoilova, B., Lau, I.J., Doondeea, J., Doondeea, B., Kennedy, A., et al. (2016). Genetically distinct leukemic stem cells in human CD34-acute myeloid leukemia are arrested at a hemopoietic precursor-like stage. J. Exp. Med. *213*, 1513–1535.
210. Estey, E.H. (2018). Acute myeloid leukemia: 2019 update on risk-stratification and management. Am. J. Hematol. *93*, 1267–1291.
211. Deschler, B., and Lübbert, M. (2006). Acute myeloid leukemia: epidemiology and etiology. Cancer *107*, 2099–2107.

212. Falini, B., Brunetti, L., and Martelli, M.P. (2021). How I diagnose and treat NPM1-mutated AML. Blood *137*, 589–599.

213. Kostopoulou, O., Tracey, C., and Delaney, B.C. (2021). Can decision support combat incompleteness and bias in routine primary care data? J. Am. Med. Inform. Assoc. *28*, 1461–1467.

214. Cabezas-Wallscheid, N., Buettner, F., Sommerkamp, P., Klimmeck, D., Ladel, L., Thalheimer, F.B., Pastor-Flores, D., Roma, L.P., Renders, S., Zeisberger, P., et al. (2017). Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. Cell *169*, 807-823.e19.

215. Söderberg, S.S., Karlsson, G., and Karlsson, S. (2009). Complex and context dependent regulation of hematopoiesis by tgf-β superfamily signaling. Ann. N. Y. Acad. Sci. *1176*, 55–69.

216. Lefort, S., and Maguer-Satta, V. (2020). Targeting BMP signaling in the bone marrow microenvironment of myeloid leukemia. Biochem. Soc. Trans. *48*, 411–418.

217. Bhatia, M., Bonnet, D., Wu, D., Murdoch, B., Wrana, J., Gallacher, L., and Dick, J.E. (1999). Bone morphogenetic proteins regulate the developmental program of human hematopoietic stem cells. J. Exp. Med. *189*, 1139–1148.

218. Raymond, A., Liu, B., Liang, H., Wei, C., Guindani, M., Lu, Y., Liang, S., St John, L.S., Molldrem, J., and Nagarajan, L. (2014). A role for BMP-induced homeobox gene MIXL1 in acute myelogenous leukemia and identification of type I BMP receptor as a potential target for therapy. Oncotarget *5*, 12675–12693.

219. Topić, I., Ikić, M., Ivčević, S., Kovačić, N., Marušić, A., Kušec, R., and Grčević, D. (2013). Bone morphogenetic proteins regulate differentiation of human promyelocytic leukemia cells. Leuk. Res. *37*, 705–712.

220. Sun, R., He, L., Lee, H., Glinka, A., Andresen, C., Hübschmann, D., Jeremias, I., Müller-Decker, K., Pabst, C., and Niehrs, C. (2021). RSPO2 inhibits BMP signaling to promote self-renewal in acute myeloid leukemia. Cell Rep. *36*.

221. Krause, C., Guzman, A., and Knaus, P. (2011). Noggin. Int. J. Biochem. Cell Biol. *43*, 478–481.

222. Miale, T.D., Stenke, L.Å.L., Lindblom, J.B., Sjögren, A.-M., Reizenstein, P.G., Udén, A.-M., and Lawson, D.L. (1982). Surface Ia-Like Expression and MLR-Stimulating Capacity of Human Leukemic Myeloblasts: Implications for Immunotherapy and Prognosis. Acta Haematol. *68*, 3–13.

223. Newman, R.A., Delia, D., Greaves, M.F., Navarrete, C., Fainboim, L., and Festenstein, H.

(1983). Differential expression of HLA-DR and DR-linked determinants on human leukemias and lymphoid cells. Eur. J. Immunol. *13*, 172–176.

224. Christopher, M.J., Petti, A.A., Rettig, M.P., Miller, C.A., Chendamarai, E., Duncavage, E.J., Klco, J.M., Helton, N.M., O'Laughlin, M., Fronick, C.C., et al. (2018). Immune Escape of Relapsed AML Cells after Allogeneic Transplantation. N. Engl. J. Med. *379*, 2330–2341.
225. Hernández-Malmierca, P. (2020). Hematopoietic stem cells are antigen presenting cells capable of inducing immunoregulatory T cell phenotypes. Submitt. to Cell.
226. Filippi, M.D., and Ghaffari, S. (2019). Mitochondria in the maintenance of hematopoietic stem cells: New perspectives and opportunities. Blood *133*, 1943–1952.
227. Hao, X., Gu, H., Chen, C., Huang, D., Zhao, Y., Xie, L., Zou, Y., Shu, H.S., Zhang, Y., He, X., et al. (2019). Metabolic Imaging Reveals a Unique Preference of Symmetric Cell Division and Homing of Leukemia-Initiating Cells in an Endosteal Niche. Cell Metab. *29*, 950-965.e6.
228. Molina, J.R., Sun, Y., Protopopova, M., Gera, S., Bandi, M., Bristow, C., McAfoos, T., Morlacchi, P., Ackroyd, J., Agip, A.-N.A., et al. (2018). An inhibitor of oxidative phosphorylation exploits cancer vulnerability. Nat. Med. *24*, 1036–1046.
229. Farge, T., Saland, E., de Toni, F., Aroua, N., Hosseini, M., Perry, R., Bosc, C., Sugita, M., Stuani, L., Fraisse, M., et al. (2017). Chemotherapy-resistant human acute myeloid leukemia cells are not enriched for leukemic stem cells but require oxidative metabolism. Cancer Discov. *7*, 716–735.
230. Baccelli, I., Gareau, Y., Lehnertz, B., Gingras, S., Spinella, J.F., Corneau, S., Mayotte, N., Girard, S., Frechette, M., Blouin-Chagnon, V., et al. (2019). Mubritinib Targets the Electron Transport Chain Complex I and Reveals the Landscape of OXPHOS Dependency in Acute Myeloid Leukemia. Cancer Cell *36*, 84-99.e8.
231. Khamis, M.M., Adamko, D.J., and El-Aneed, A. (2017). Mass spectrometric based approaches in urine metabolomics and biomarker discovery. Mass Spectrom. Rev. *36*, 115–134.
232. Zhang, Y., Lu, H., Shen, Y., Chen, R., Fang, P., Yu, H., Wang, C., and Jia, W. (2015). Analysis of reproducibility and variability from a frozen sample aliquotter by metabolomics analysis. Biopreserv. Biobank. *13*, 20–24.
233. Breier, M., Wahl, S., Prehn, C., Fugmann, M., Ferrari, U., Weise, M., Banning, F., Seissler, J., Grallert, H., Adamski, J., et al. (2014). Targeted metabolomics identifies reliable and stable metabolites in human serum and plasma samples. PLoS One *9*, e89728.

234. Yu, V.W.C., Yusuf, R.Z., Oki, T., Wu, J., Saez, B., Wang, X., Cook, C., Baryawno, N., Ziller, M.J., Lee, E., et al. (2017). Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. Cell *168*, 944–945.

235. Desai, P., Mencia-Trinchant, N., Savenkov, O., Simon, M.S., Cheang, G., Lee, S., Samuel, M., Ritchie, E.K., Guzman, M.L., Ballman, K. V, et al. (2018). Somatic mutations precede acute myeloid leukemia years before diagnosis. Nat. Med. *24*, 1015–1023.

236. Loberg, M.A., Bell, R.K., Goodwin, L.O., Eudy, E., Miles, L.A., SanMiguel, J.M., Young, K., Bergstrom, D.E., Levine, R.L., Schneider, R.K., et al. (2019). Sequentially inducible mouse models reveal that Npm1 mutation causes malignant transformation of Dnmt3a-mutant clonal hematopoiesis. Leukemia *33*, 1635–1649.

237. Guryanova, O.A., Shank, K., Spitzer, B., Luciani, L., Koche, R.P., Garrett-Bakelman, F.E., Ganzel, C., Durham, B.H., Mohanty, A., Hoermann, G., et al. (2016). DNMT3A mutations promote anthracycline resistance in acute myeloid leukemia via impaired nucleosome remodeling. Nat. Med. *22*, 1488–1495.

238. Venugopal, K., Feng, Y., Shabashvili, D., and Guryanova, O.A. (2021). Alterations to DNMT3A in Hematologic Malignancies. Cancer Res. *81*, 254–263.

239. Jyoti Nangalia, Francesca L. Nice, David C. Wedge, Anna L Godfrey, Jacob Grinfeld, Clare Thakker, Charlie E. Massie, Joanna Baxter, David Sewell, Yvonne Silber, et al. (2015). DNMT3A mutations occur early or late in patients with myeloproliferative neoplasms and mutation order influences phenotype. Haematologica *100*, e438–e442.

240. Izzo, F., Lee, S.C., Poran, A., Chaligne, R., Gaiti, F., Gross, B., Murali, R.R., Deochand, S.D., Ang, C., Jones, P.W., et al. (2020). DNA methylation disruption reshapes the hematopoietic differentiation landscape. Nat. Genet. *52*, 378–387.

241. Kim, E., Goren, A., and Ast, G. (2008). Insights into the connection between cancer and alternative splicing. Trends Genet. *24*, 7–10.

242. Gareth James Trevor Hastie, Robert Tibshirani, D.W. (2013). An introduction to statistical learning : with applications in R (New York : Springer).

243. Modarres, P., Mohamadi Farsani, F., Nekouie, A.A., and Vallian, S. (2021). Meta-analysis of gene signatures and key pathways indicates suppression of JNK pathway as a regulator of chemo-resistance in AML. Sci. Rep. *11*, 1–16.

244. Martineau, E., Tea, I., Loaëc, G., Giraudeau, P., and Akoka, S. (2011). Strategy for choosing extraction procedures for NMR-based metabolomic analysis of mammalian cells. Anal. Bioanal. Chem. *401*, 2133.

245. Dietmair, S., Timmins, N.E., Gray, P.P., Nielsen, L.K., and Krömer, J.O. (2010). Towards

quantitative metabolomics of mammalian cells: Development of a metabolite extraction protocol. Anal. Biochem. *404*, 155–164.

246. Wilson, Z.E., Rostami-Hodjegan, A., Burn, J.L., Tooley, A., Boyle, J., Ellis, S.W., and Tucker, G.T. (2003). Inter-individual variability in levels of human microsomal protein and hepatocellularity per gram of liver. Br. J. Clin. Pharmacol. *56*, 433–440.
247. Sohlenius-Sternbeck, A.K. (2006). Determination of the hepatocellularity number for human, dog, rabbit, rat and mouse livers from protein concentration measurements. Toxicol. Vitr. *20*, 1582–1586.
248. Calderón, C., Sanwald, C., Schlotterbeck, J., Drotleff, B., and Lämmerhofer, M. (2019). Comparison of simple monophasic versus classical biphasic extraction protocols for comprehensive UHPLC-MS/MS lipidomic analysis of Hela cells. Anal. Chim. Acta *1048*, 66–74.
249. Gegner, H.M., Mechtel, N., Heidenreich, E., Wirth, A., Cortizo, F.G., Bennewitz, K., Fleming, T., Andresen, C., Freichel, M., Teleman, A.A., et al. (2022). Deep Metabolic Profiling Assessment of Tissue Extraction Protocols for Three Model Organisms. Front. Chem. *10*.
250. Fu, X., Calderón, C., Harm, T., Gawaz, M., and Lämmerhofer, M. (2022). Advanced unified monophasic lipid extraction protocol with wide coverage on the polarity scale optimized for large-scale untargeted clinical lipidomics analysis of platelets. Anal. Chim. Acta *1221*, 340155.
251. Andresen, C., Boch, T., Gegner, H.M., Mechtel, N., Narr, A., Birgin, E., Rasbach, E., Rahbari, N., Trumpp, A., Poschet, G., et al. (2022). Comparison of extraction methods for intracellular metabolomics. Front. Mol. Biosci. *9*.
252. Kim-Wanner, S.-Z., Luxembourg, B., Schmidt, A.H., Schäfer, R., Möller, N., Herbert, E., Poppe, C., Hümmer, C., Bunos, M., Seifried, E., et al. (2020). Introduction of principles of blood management to healthy donor bone marrow harvesting. Vox Sang. *115*, 802–812.
253. Network, T.C.G.A.R. (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N. Engl. J. Med. *368*, 2059–2074.
254. Goldman, M.J., Craft, B., Hastie, M., Mcdade, F., Banerjee, A., Luo, Y., Rogers, D., Brooks, A.N., Haussler, D., Cruz, S., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. Nat. Biotechnol. *38*, 675–678.
255. Fuhrer, T., Heer, D., Begemann, B., and Zamboni, N. (2011). High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass

spectrometry. Anal. Chem. *83*, 7074–7080.

256. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

257. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

258. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

259. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. Bioinformatics *31*, 2032–2034.

260. Liao, Y., Smyth, G.K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

261. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 1–21.

262. Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. *36*, 421–427.

263. Zhu, A., Ibrahim, J.G., and Love, M.I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics *35*, 2084–2092.

264. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omi. A J. Integr. Biol. *16*, 284–287.

265. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

266. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.e29.

267. Mullen, K.M., and von Stokkum, I.H.M. (2012). nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS). R package version 1.4.

268. Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H., and Califano, A.

(2016). Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat. Genet. *48*, 838–847.

269. Holland, C.H., Szalai, B., and Saez-Rodriguez, J. (2020). Transfer of regulatory knowledge from human to mouse for functional genomics analysis. Biochim. Biophys. acta. Gene Regul. Mech. *1863*, 194431.
270. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. *29*, 1363–1375.
271. Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat. Commun. *9*.
272. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.
273. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.
274. Therneau, T.M. (2022). A Package for Survival Analysis in R. R package version 3.3-1.
275. Therneau, T.M., and Grambsch, P.M. (2000). Modeling survival data : extending the Cox model (Springer).
276. Alboukadel, K., Kosinski, M., Biecek, P., and Scheipl, F. (2021). survminer: Drawing Survival Curves using “ggplot2.” R package version 0.4.9.
277. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.
278. Sheffield, N.C., and Bock, C. (2016). LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics *32*, 587–589.
279. Sánchez-Castillo, M., Ruau, D., Wilkinson, A.C., Ng, F.S.L., Hannah, R., Diamanti, E., Lombard, P., Wilson, N.K., and Gottgens, B. (2015). CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. Nucleic Acids Res. *43*, D1117-23.
280. Guangchuang Yu (2019). enrichplot: Visualization of Functional Enrichment Result. R package version 1.4.0.
281. Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics *31*, 3718–3720.

282. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, 1–13.

283. van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., and van der Werf, M.J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics *7*, 142.