MODELING AGE-DEPENDENT GENE EXPRESSION VARIABILITY IN ACUTE MYELOID LEUKEMIA USING A LINEAR MODEL

By

Raeuf Roushangar

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry and Molecular Biology—Doctor of Philosophy
Quantitative Biology—Dual Major

2018

ABSTRACT

MODELING AGE-DEPENDENT GENE EXPRESSION VARIABILITY IN ACUTE MYELOID LEUKEMIA USING A LINEAR MODEL

By

Raeuf Roushangar

In 2018 alone, an estimated 20,000 new acute myeloid leukemia (AML) patients were diagnosed, in the United States, and over 10,000 of them are expected to die from the disease. Although AML can occur in people of all ages, AML is primarily diagnosed among the elderly (median 68 years old at diagnosis) and its age-specific incidence and prevalence increases exponentially after 50 years of age. Prognoses have significantly improved for younger patients, but in patients older than 60 years old, prognoses remain grim: with current treatments, as much as 70% of patients will die within a year of diagnosis. Reassessment of early diagnosis and treatment approaches therefore should be considered, since relapse after complete remission is still the main obstacle. In this study, we conducted stratified computational meta-analysis of 2,213 AML patients compared to 548 healthy individuals, using curated publicly available data. We carried out analysis of variance of normalized batch corrected data, including considerations for disease, age, tissue and sex. We identified 964 differentially expressed unique genes genes and 4 associated significant pathways involved in AML. Additionally, we have identified 69 sex- and 372 age-related gene expression signatures relevant to AML. Finally, we used a machine learning model (KNN model) to classify AML patients compared to healthy individuals with > 90% achieved accuracy. Overall our findings provide a new reanalysis of public datasets, that enabled the identification of potential new gene sets relevant to

AML that can potentially be used in future experiments and possible stratified disease diagnostics.

I dedicate this book to my mother for detecting her entire life to raise me with unconditional love as a single mother. I dedicate this work to this wonderful country for opening its doors and gave me a shot. It’s a true privilege to wake up every day and be able to think freely, which I couldn't do for so many years, and do scientific research with a shot to contribute to science and humanity!

ACKNOWLEDGMENTS

First and foremost, I would like to thank my PhD advisor, Dr. George I. Mias, who not only welcomed me to his lab and mentored me with kindness and calm, but also both challenged me and helped me mature as a scientist. Dr. Mias gave all necessary resources and created a great environment for me to do scientific research freely. He gave me space to express and discuss my views without worrying, this is very monumental to me. He supported me and let me reincarnate myself into a form that I have always wanted, which allowed me to determine who I am and what I wanted to become as a scientist. I want to deeply thank him for teaching me how to be a scientist who strive for depth, aware of fundamental assumptions, and to always be objective. I am forever in his debt.

I would also like thank the people in the lab that made the last four years fun and possible. Dr. Vikas V. Singh and Lavida R. K. Brooks didn't just edit my scientific thinking through discussions and interpret results, but they became my friends which I deeply value. I am also deeply indebted to my close friends, Timothy Stachelski and Qingfeng Zeng, who supported me through the years, edited my thoughts, and were always there for me when I needed them. I appreciate their friendship very much.

I am also extremely thankful for having Dr. Ronald Henry, Dr. A. Daniel Jones, Dr. Carlo Piermarocchi, Dr. Curtis Wilkerson, and Dr. Timothy Zacharewski as my guiding committee. My time as a PhD student would have been dramatically different if not for their guide and constant support. I want to also thank Dr. Thomas D. Sharkey Dr. Jon

Kaguni and Mrs. Jessica Lawrence for guiding me through my graduate program and helping on many occasions.

I am also thankful to Dr. Brian Schutte for giving me the opportunity and resources to conduct research in his lab during my undergraduate time at Michigan State University. I am also deeply indebted to Dr. Youssef A. Kousa for dedicating his time to mentor me during my 3 years at Dr. Schutte's lab, which helped me grow as a future scientist and develop my own individual creative spirit.

Finally, I want to deeply thank Dr. Pamela J. Fraker and Dr. John LaPres for giving me a in helping me to be admitted into the PhD program at Michigan State University. I would also like to thank The Paul & Daisy Soros Fellowships for supporting this work, which allowed me to tackle research problems without any financial burdens.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

AML.....acute myeloid leukemia

WHO……..World Health Organization

AML-RGA..…….aml with recurrent genetic abnormalities

t(8;21)……..……t(8;21)(q22;q22.1);RUNX1-RUNX1T1

RUNX1……Runt Related Transcription Factor 1

CBFB……..Core-Binding Factor Beta Subunit

inv(16) or t(16;16)…….....inv(16)(p13.1q22) or t(16;16)(p13.1;q22);CBFB-MYH11

MYH11……...……Myosin Heavy Chain 11

APL…….……..…..acute promyelocytic leukemia

PML.……...Promyelocyte Leukemia

RARA……Retinoic Acid Receptor Alpha

t(9;11)……..…….t(9;11)(p21.3;q23.3);MLLT3-KMT2A

KMT2A……...Lysine Methyltransferase 2A

MLLT3……...MLLT3, Super Elongation Complex Subunit

AMGL……..…...…..….acute myelogenous leukemia

t(6;9)…….t(6;9)(p23;q34.1);DEK-NUP214

DEK……DEK Proto-Oncogene

NUP214……...Nucleoporin 214

CR……...complete remission

inv(3) or t(3;3)..……...…inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2, MECOM

GATA2……..GATA Binding Protein 2

MECOM……..MDS1 And EVI1 Complex Locus

AMKL…….….acute megakaryoblastic leukemia

t(1;22)……..t(1;22)(p13.3;q13.3);RBM15-MKL1

RBM15…….RNA Binding Motif Protein 15

MKL1……..Megakaryoblastic Leukemia (Translocation) 1

BCR……...BCR, RhoGEF And GTPase Activating Protein

ABL1……...ABL Proto-Oncogene 1, Non-Receptor Tyrosine Kinase

CML…….chronic myeloid leukemia

ALL……acute lymphoblastic leukemia

NPM1……..…...Nucleophosmin 1

CEBPA……CCAAT Enhancer Binding Protein Alpha

bi-CEBPA..……..…….....biallelic mutations of CEBPA

AML-MRC…….....aml with myelodysplasia-related changes

OS…….overall survival

MDS……myelodysplastic syndrome

MDS/MPN……..myelodysplastic/myeloproliferative neoplasms

t-MN……..therapy-related myeloid neoplasms

AML-NOS…….aml not otherwise specified

GUI……...graphical user interface

GEO……Gene Expression Omnibus

BM……bone marrow

PB……peripheral blood

KNN…….k-nearest neighbor

LR…….logistic regression

PCA…….principal component analysis

DGE……...differential gene expression

DEPS……...differentially expressed probe sets

ANOVA……...Analysis of Variance

HSD……Honestly Significant Difference

WT1……...Wilms tumor 1

CRISP3……...cysteine-rich secretory protein 3

KEGG…….Kyoto Encyclopedia of Genes and Genomes

GO…….……Gene Ontology

DAVID……...Database for Annotation, Visualization and Integrated Discovery

JUP……junction plakoglobin

CCNA1……...cyclin A1

FLT3……fms-related tyrosine kinase 3

PIK3R1……..phosphoinositide-3-kinase, regulatory subunit 1 (alpha)

CD14…….CD14 molecule

CEBPE……CCAAT/enhancer binding protein (C/EBP), epsilon

DDX3Y…….DEAD-Box Helicase 3 Y-Linked

EIF1AY……Eukaryotic Translation Initiation Factor 1A Y-Linked

KDM5D……...Lysine Demethylase 5D

RPS4Y1……...Ribosomal Protein S4 Y-Linked 1

XIST……..X Inactive Specific Transcript

TSIX……...TSIX Transcript, XIST Antisense RNA

PRKX…….Protein Kinase X-Linked

HOX……homeobox genes

ORM1……Orosomucoid 1

NCBI……..National Center for Biotechnology Information

RMA……Robust Multi-Array Average

PM……perfect match

DGIdb……..The Drug Gene Interaction Database

Chapter 1 -

Acute Myeloid Leukemia

Historical

Cancer was first identified and described in Egypt, where evidence from ancient Egyptian mummies and manuscripts date back more than 3,500 years [1]. It was the Greek physician Hippocrates (460-370 BC) however, who coined the word, "cancer" (καρκίνος in Greek) [1]. According to Gordon Piller [2], blood malignancies were hard to diagnose since microscopic examination of the blood was not possible until Robert Hooke published work on microscopy in 1665 [3]. In 1674, Anton van Leeuwenhoek was the first to describe human red blood cells, and in 1749, white blood cells (including lymphocytes) were described by Joseph Lieutaud [2,4].

In 1845, John Hughes Bennett, then pathologist at the Royal Infirmary of Edinburgh, carried out the post mortem of a patient and reported that the patient's blood was affected throughout his system [5]. Around the same time, other cases with blood abnormality were reported by Rudolf Virchow in Berlin and Henry Fuller in London [6,7]. The findings of Bennett, Virchow, and Fuller led to the recognition of leukemia as a distinct disease [2]. The earliest recorded case of acute leukemia, a form of leukemia, took place in 1857 when the German pathologist Nikolaus Friedreich observed mass leukocytes formed in his 46-year-old patient's thorax 6 weeks before her death [8].

In 1880, Paul Ehrlich developed staining methods to stain and trace blood cells – his work led to the classification of myeloid and lymphoid leukemia subtypes [2,9]. One of the earliest recorded epidemiological studies was of 154 cases of leukemia, which took place in 1879 when W. R. Gowers and others speculated that the disease might be due to

exposure to malaria [10]. In 1894, the work of Dr. Richard Cabot, then a physician in Boston, was vital to the recognition of acute leukemia where he published his work on 34 patients that had an average survival of 4.5 weeks after their diagnosis [11]. And In 1909 Dr. Robert J. M. Buchanan clinically described acute myeloid leukemia (AML), its onset and rapid progression [2,12].

AML statistics and incidences

Each year, cancer affects millions of people in the United States (US) and around the world [13-15]. Within the US, cancer is the second leading cause of death after heart disease with 1,735,350 new cases and 609,640 deaths projected for 2018 [14]. Leukemia is a cancer of the blood and is currently the 9th most common type of cancer and the 6th leading cause of death in males and 7th in females in the US [14]. Myeloid leukemia is the most common type of leukemia, and AML accounts for 70% of myeloid leukemia and nearly 80% of acute leukemia cases, making it the most common form of both myeloid and acute leukemia [14,16,17]. The number of new AML cases is increasing each year – in 2018 alone, there have been an estimated 60,300 new diagnosed leukemia patients. About 20,000 of these are AML cases, over 10,000 of which will die from the disease [18]. In fact, AML has the highest mortality rate of all leukemia related disease [19].

AML characteristics

AML is a blood cancer that best described as several heterogeneous diseases with many complex genetic abnormalities. Specifically, AML is a multifactorial cancer of the myeloid cell lineage of the hematopoietic system that begins in the bone marrow. AML is

characterized by terminal differentiation of normal blood cells and excessive proliferation and release of abnormally differentiated myeloid cells (leukemia cells) at various stages of myeloid hematopoiesis [20]. This faster than normal and uncontrolled growth leads to abnormal accumulation and buildup of leukemic cells in the bone marrow and peripheral blood, frequently resulting in suppression of healthy myeloid precursors of the hematopoietic system and hematopoiesis insufficiency [20].

AML classification

According to the 2016 World Health Organization (WHO) newly revised myeloid neoplasms and acute leukemia classification system, there are a number of major disease categories of AML and many subtypes (Table 1) [21]. This classification system is based on factors that affect AML prognosis, including cytogenetic abnormalities, molecular genetic alterations, morphologic features, immunophenotypic, and biological and clinical information [16,21]. The major categories of AML classification are, 1) 'AML with recurrent genetic abnormalities', 2) 'AML with myelodysplasia-related changes', 3) 'Therapy-related myeloid neoplasms', and 4) 'AML not otherwise specified', described below.

AML with recurrent genetic abnormalities

Hereafter abbreviated AML-RGA. Chromosomal abnormalities including deletions, duplications, translocations, inversions, and gene fusion occur frequently in AML [22]. AML-RGA encompasses a number of different AML subgroups with specific distinctive chromosomal abnormalities that include we list here, and further discuss below:

- 'AML with t(8;21)(q22;q22.1);RUNX1-RUNX1T1'

- ‘AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22);CBFB-MYH11’
- ‘Acute promyelocytic leukemia with PML-RARA’
- ‘AML with t(9;11)(p21.3;q23.3);MLLT3-KMT2A’
- ‘Acute myelogenous leukemia with t(6;9)(p23;q34.1);DEK-NUP214’
- ‘AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2, MECOM’
- ‘Acute megakaryoblastic leukemia with t(1;22)(p13.3;q13.3);RBM15-MKL1’
- ‘AML with BCR-ABL1’
- ‘AML with mutated NPM1’
- ‘AML with biallelic mutations of CEBPA’
- and ‘AML with mutated RUNX1’ [16,21].

- **AML with t(8;21)(q22;q22.1);RUNX1-RUNX1T1**

Hereafter abbreviated AML with t(8;21). Translocation in chromosomes 8 and 21, t(8;21), is one of the most common AML chromosomal abnormalities and is associated with 12% of all AML cases [23]. In 1973, Dr. Janet Rowley was first to discover the translocation and breaks at q22;q22 in chromosomes 8 and 21 in a female patient with acute leukemia [24]. In early 1990, the location of RUNX1 and RUNX1T1 were identified to be at the translocation site [23]. In 1993 Miyoshi et al. (1993) [25] reported that the t(8;21) translocation in AML results in the RUNX1-RUNX1T1 fusion protein. RUNX1 is a gene encoding DNA-binding transcription factor that binds to DNA using its runt-homology domain and interacts with CBFB, a common heterodimeric partner [26].

- **AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22);CBFB-MYH11**

Hereafter abbreviated AML with inv(16) or t(16;16). The inversion and/or translocation of chromosome 16, inv(16) or t(16;16), is among the most frequently observed chromosomal abnormalities found in AML and is detected in about 16% of AML cases [27]. In 1983, Le Beau et al. (1983) [28] were first to report inv(16) in leukemic cells from newly diagnosed AML patients with abnormal bone marrow. In 1993 Dr. Paul Liu identified the two genes, CBFB and MYH11, located at the inversion breakpoints, that resulted in chimeric mRNA formation, which generates CBFB-MYH1 a fusion protein product resulting from inv(16) in AML [29,30]. CBFB, located at 16q22, encodes the beta subunit of the core binding transcription factor, whereas MYH11, located at 16p13.1, encodes the smooth muscle myosin heavy chain 11.

RUNX1 and CBFB are both crucial to transcriptional regulation of healthy hematopoiesis development [26,31]. Chromosomal abnormalities and mutations in RUNX1 and CBFB result in terminal differentiation of healthy myeloid cells and uncontrolled proliferation of leukemia cells at various stages of hematopoiesis, which ultimately leads to hematological malignancies [26,32]. AML with t(8;21) and AML with inv(16) or t(16;16) are classified as core binding factor (CBF) AML and together they account for approximately 20% of all adult AML cases [31]. These AML subtypes are commonly associated with favorable prognosis and response to conventional therapy [33].

- **Acute promyelocytic leukemia with PML-RARA**

Acute promyelocytic leukemia (APL) is a subtype of AML that has distinct and clear biological features. APL accounts for approximately 10% of all AML cases [34,35]. In 1957 Dr. Leif Hillestad was first to identify APL and characterized its clinical features [36]. In 1977 Rowley et al. (1977) [37] identified the APL cytogenetic signature as the reciprocal translocation between chromosome 15 and 17, t(15;17), which results in fusion of the PML and RARA genes [38]. RARA is involved in transcriptional regulation, gene expression, and various other biological processes, including its function as a ligand-dependent receptor for retinoic acid binding [39-41]. The PML-RARA fusion protein represses the retinoic acid downstream response, and targets gene expression, resulting in abnormal and uncontrolled cell proliferation and suppression of normal cellular process [42,43].

- **AML with t(9;11)(p21.3;q23.3);MLLT3-KMT2A**

Hereafter abbreviated AML with t(9;11). Chromosomal abnormalities that lead to the translocation between chromosome 9 and 11, t(9;11), result in fusion of the KMT2A gene (also known as MLL, MLL1, ALL1) with the MLLT3 gene (also known as AF9). AML patients with MLLT3-KMT2A fusion as a result of translocation (9;11) usually have short survival rate, frequent disease relapse, and poor clinical outcome [44-47]. KMT2A gene rearrangement has been reported in approximately 10% of all acute leukemia cases [48]. KMT2A, located at 11q23, encodes a transcription factor that is involved in gene expression regulation essential to chromatin remodeling, development, and hematopoiesis [49].

- **Acute myelogenous leukemia with t(6;9)(p23;q34.1);DEK-NUP214**

Hereafter abbreviated AMGL with t(6;9). AMGL with chromosomal aberration t(6;9)(p23;q34.1) is a rare form of AML and is observed in about 0.5% to 4% of all AML patients [50]. In 1976, Dr. Janet Rowley and Dr. David Potter studied bone marrow samples obtained from 50 adult patients and reported the translocation between chromosome 6 and 9 in AMGL [51]. The translocation between chromosome 6 and 9, t(6;9), results in fusion of the DEK gene (located at 6p23) with the NUP214 gene (located at 9q34) [50,52]. AML patients with the chimeric DEK-NUP214 fusion gene have poor prognosis and only 50% achieve complete remission (CR) with conventional chemotherapy [52].

- **AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2, MECOM**

Hereafter abbreviated AML with inv(3) or t(3;3). Chromosomal abnormalities that lead to inversion and/or translocation in the chromosome 3 long arm have been detected in approximately 1% to 2% of all AML cases [53]. In particular, inv(3) abnormalities have been the most frequently observed chromosomal abnormalities in this group and are associated with poor prognosis, treatment response, and median survival rate of less than 1 year [53-56]. The GATA2 gene encodes for transcription factor GATA binding protein 2, an important regulator of hematopoietic cell differentiation [57], whereas the MECOM gene (also known as EV1, MDS1) encodes for the transcriptional regulator MDS1 and EVI1 complex locus, which is involved in cell differentiation essential to development and hematopoiesis [58].

Recently, Gröschel et al. (2014) [59] and Yamazaki et al. (2014) [60] revealed inv(3) biology in AML: they discovered that in inv(3), the GATA2 enhancer is repositioned from 3q21 to be in close proximity with the MECOM gene at 3q26. This rearrangement in turn activates MECOM gene expression and causes GATA2 haploinsufficiency at its original location, which ultimately leads to leukemogenesis [59,60].

- **Acute megakaryoblastic leukemia with t(1;22)(p13.3;q13.3);RBM15-MKL1**

Hereafter abbreviated AMKL with t(1;22). AMKL is a rare hematologic malignant disease that is detected in <1% of all AML patients [61]. AMKL is closely associated (high incidence) with infants and young children [62-64]. In 1991, the translocation between chromosome 1 and 22, t(1;22), was first reported as the principal nonrandom cytogenetic signature in infants with AMKL [65,66]. In 2001 Ma et al. (2001) [67] reported that this chromosomal translocation fuses two novel genes, RBM15 gene (also known as OTT), located at 1p13, and MKL1, located at 22q13, which generates the RBM15-MKL1 chimeric protein product.

The RBM15 gene encodes three RNA-recognition motifs involved in modulating Hox homeotic function, which regulates the Ras/MAP kinase signaling essential to cell differentiation and proliferation [68]. The MKL1 gene encodes an SAP DNA binding domain that is involved in transcription regulation, chromatin remodeling, and extracellular signaling pathways [67,69,70]. AMKL patients with RBM15-MKL1 fusion as a result of the t(1;22) translocation have poor prognosis and clinical course with less than 1 year survival time from diagnosis [63].

- **AML with BCR-ABL1**

Chromosomal abnormalities leading to the translocation between chromosome 9 and 22 result in the BCR-ABL1 fusion gene, commonly referred to as the Philadelphia chromosome. It is most frequently associated with chronic myeloid leukemia (CML) and acute lymphoblastic leukemia (ALL). AML with BCR-ABL1 accounts for 0.5% to 3% of all AML cases [71-74]. Because of recent improvement in the reliability and standardization of diagnosis for this rare disease, AML with BCR-ABL1 was recently added as a provisional entity in the 2016 WHO newly revised myeloid neoplasms and acute leukemia classification system [21].

The ABL1 gene encodes the ABL protooncogene non-receptor tyrosine kinase protein involved in cell division and apoptosis [75]. The function of the BCR gene product is complex, however, Duejmann et al. (1991) [76] and Maru et al. (1991) [77] showed the participation of BCR in eukaryotic intracellular signaling via phosphorylation and GTP-binding [78]. Since BCR-ABL1 fusion affects the regulation of hematopoietic cells [78], AML patients with the aberrant BCR-ABL1 fusion gene have unfavorable prognosis and are among AML poor risk group [79].

- **AML with mutated NPM1**

Mutations in the NPM1 gene are the most common and frequent mutations found in AML patients – they are detected in approximately 30% and 60% of all AML patients and AML patients with normal karyotype, respectively [80,81]. The NPM1 gene encodes the nucleophosmin 1 protein, which is a member of the nucleophosmin/nucleoplasmin

proteins family [82]. Under normal conditions, the NPM1 protein is mainly restricted to the nucleolus, but shuttles between the nucleus, where it modulates pre-ribosomal protein nuclear export, and the cytoplasm, where it regulates centrosome duplication, during the cell cycle [81,83].

In 2005 Brunangelo et al (2005) [80] examined bone marrow specimens from 591 primary AML patients and found that 208 (35.2%) of the 591 AML patients have NPM1 gene mutations and cytoplasmic dislocation of the NPM1 protein, and suggested that NPM1 gene mutations that cause changes in the NPM1 protein are responsible for the translocation of the NPM1 protein from the nucleus to the cytosol. Cytoplasmic dislocation of the NPM1 protein as a result of NPM1 genetic mutations is thought to play a major role in leukemogenesis [81].

NPM1 is important in many cellular processes, including DNA repair and cell survival [84], ribosome biogenesis [85], chromatin remodeling [86], protein chaperoning [87], and regulation of the ARF–tumor suppressors p53 pathway [80,88,89]. AML patients with NPM1 gene mutations, with normal karyotype and absence of Fms related tyrosine kinase 3 internal tandem duplication (FLT3 ITD) mutations, continue to be associated with favorable AML prognosis, response to conventional therapy, achieve CR, and are among an AML favorable risk group [80,81,90].

- **AML with biallelic mutations of CEBPA**

Hereafter abbreviated AML with bi-CEBPA. The presence of mutations in the CEBPA gene ranges from 10% to 15% of all AML cases, and these are closely associated with AML patients who have cytogenetically normal karyotype [91,92]. The CEBPA gene is a transcription factor involved in and is upregulated during progenitor cell differentiation and proliferation [93,94]. Mutations in the CEBPA gene lead to terminal or abnormal cell differentiation, and ultimately to leukemogenesis [94,95].

The CEBPA gene has two hotspots where mutations cluster: the N-terminal, where frame-shift (insertions/deletions) mutations affect the CEBPA transactivation domains, and the C-terminal, where in-frame mutations (insertions/deletions) in the DNA-binding motif affect protein dimerization and DNA binding [95-97]. AML patients with biallelic mutations of CEBPA (bi-CEBPA) – mutations on both CEBPA alleles – have mutations on both hotspots: Frame-shift and in-frame mutations on the N- and C- terminus, respectively [98]. Furthermore, only AML with bi-CEBPA patients are uniquely associated with favorable clinical outcome and improved survival, but AML patients with single CEBPA mutations or wild-type CEBPA are not [99,100].

- **AML with mutated RUNX1**

RUNX1 is a transcription factor that is expressed in healthy hematopoietic cells essential to the hematopoietic system. In adults, mutations in the RUNX1 gene lead to AML [101,102]. AML with mutated RUNX1 accounts for approximately 10% of all AML cases and is associated with newly-diagnosed (*de novo*) AML patients [103-105]. Gaidzik et al. (2016)

investigated the RUNX1 gene mutation frequency and prognosis in 2439 *de novo* AML patients and reported that 245 (10%) of the 2439 AML patients had RUNX1 gene mutations with almost no chromosomal abnormalities [105]. The disease was recently added as a provisional entity in the 2016 WHO newly revised myeloid neoplasms and acute leukemia classification system, since AML patients with mutated RUNX1 have distinct biological features including poor disease prognosis with worse overall survival (OS) compared to other AML types [21].

AML with myelodysplasia-related changes

AML with myelodysplasia-related changes (AML-MRC) is a heterogeneous disease that is closely associated with myelodysplastic syndromes (MDS), myelodysplastic/myeloproliferative neoplasms (MDS/MPN), poor prognosis, and elderly patients [21]. AML-MRC cases account for approximately 20% to 30% of all AML cases and are among an AML poor risk group [21,106]. Cell morphology, cytogenetic abnormalities, and clinical features are important factors for AML-MRC prognosis [21].

According to the 2016 WHO newly-revised 'myeloid neoplasms and acute leukemia' classification system, AML-MRC classification criteria include: 1) presence of 50% or more multilineage dysplasia (two or more cell lines) with no presence of NPM1 gene mutations or biallelic mutated CEBPA and 2) presence of at least 20% blast cells in the bone marrow or peripheral blood, and/or 3) previous history of MDS or MDS/MPN, or presence of MDS or MDS/MPN related cytogenetic abnormalities (Table 2) -- excluding abnormalities associated with NPM1 gene mutations or biallelic mutated CEBPA

(del(9q)), or related to AML-RGA, or to prior cytotoxic therapy used for unrelated disease [107-111].

Therapy-related myeloid neoplasms

Therapy-related myeloid neoplasms (t-MN) encompass a number of therapy-related malignant diseases (therapy-related AML (t-AML), MDS (t-MDS), and MDS/MPN (t-MDS/MPN)) occurring as a function of prior cytotoxic therapy (radiation, chemo, or both) where the therapy used is independent of diseases (malignant or not) [112].

There are two major clinical subtypes of t-MD [113]. The first subtype is approximately 70% of all t-MN and is associated with chromosomal abnormalities that result in the deletion of part of chromosome 5 (del(5q)) and/or part or all of chromosome 7 (del(75q)/-7) [114]. Furthermore, this subtype is associated with patients who received prior alkylating/radiation therapy, have poor prognosis, and initially diagnosed with MDS that progresses to AML [114]. Finally, the median survival of patients with this subtype is 8 months. The second subtype of t-MN is associated with chromosomal abnormalities that results in the translocation of the KMT2A gene (located at 11q23) or the RUNX1 gene (located 21q22) [113,114]. Moreover, this subtype is associated with patients who received prior topoisomerase II inhibitors therapy, diagnosed as leukemia with absence of MDS, and have favorable clinical outcomes with standard treatment [114].

AML not otherwise specified

AML not otherwise specified (AML-NOS) encompasses a number of heterogenous AML subtypes (not classified in any of the other three AML categories (AML-RGA, AML-MRC, and t-MN)) [21]. According to the WHO, AML-NOS subtypes include, 'AML with minimal differentiation', 'AML without maturation', 'AML with maturation', 'acute myelomonocytic leukemia', 'acute monoblastic/monocytic leukemia', 'pure erythroid leukemia', 'acute megakaryoblastic leukemia', 'acute basophilic leukemia', and 'acute panmyelosis with myelofibrosis' [21,115]. Since AML-NOS patients have intermediate prognosis and lacks consistent diagnostic criteria such as clinical features or cytogenetic abnormalities, AML-NOS patients are often classified based on cell morphology, pancytopenia and/or bone marrow dysfunction [115].

AML complex genetic abnormalities suggests that AML evolves over time

AML is typically diagnosed through microscopic, cytogenetics, and molecular genetic analyses of patients' blood and bone marrow samples. Microscopic examination is used to detect distinctive features (e.g. Auer rods) in cell morphology, cytogenetic analysis to identify chromosomal structural aberrations (e.g., t(8;21), inv(16) or t(16;16), t(9;11)), and molecular genetic analysis to identify mutations in genes frequently mutated in AML (e.g., NPM1, RUNX1, FLT3) [116-118].

Cytogenetic and molecular genetic analyses are used to identify prognosis markers that can be used to classify AML patients into three risk categories: favorable, intermediate, and unfavorable. The largest group of AML patients (almost 50%) however, present

normal karyotype and lack genetic abnormalities [117-121]. These patients are classified as intermediate risk, and often have heterogeneous clinical outcome with standard therapy with risk of AML relapse [122,123]. Further complicating, AML has multiple driver mutations and competing clones that evolve over time, making it a very dynamic disease [124-126].

In 2013 The Cancer Genome Atlas (TCGA) et al. (2013) [124] published a study entitled "Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia", and generated a catalogue of 23 genes that are significantly mutated in AML (Fig. 1). Building upon previous findings, these discoveries are expanding our knowledge of AML as well as revealing how biologically complex AML is, but the extent of their utility in disease prognosis, clinical practice, and patients' clinical outcomes are as yet unclear [127].

AML standard treatments

Briefly, AML standard treatments consist of two phases: Remission induction therapy to eradicate as many leukemia cells as possible and to produce a CR in the bone marrow, followed by an intensive consolidation phase (post-remission) to prevent AML relapse [22,116]. Generally, treatments employ a 7 + 3 regiment that consists of two chemo drugs: A 7-day continuous infusion of standard-dose cytarabine, and 3 days of anthracycline, daunorubicin or idarubicin, followed by the consolidation phase if CR is achieved (otherwise a second induction course can be considered) [116,128].

Cytarabine, typically used in AML treatment, inhibits polymerase activity and damages DNA during the cell cycle, while daunorubicin, another chemotherapy medication, intercalates between base pairs of DNA/RNA strands, which prevents DNA replication and inhibits the enzyme topoisomerase II from relaxing supercoiled DNA. With current standard treatments, 40% of younger patients have 5-year OS. The majority of AML cases are elder patients (> 60 years) and they have no standard treatment option, which is reflected in a much lower 5-year OS: 10% to 20% [22,116]. This is because older AML patients have poor clinical outcomes, decreased sensitivity to chemotherapy, and adverse cytogenetic abnormalities due to their lack of tolerability to the ideal dose of chemotherapy [22,129-134]. While other treatments, including intensive chemotherapy and immunotherapies for younger patients, and hematopoietic stem cell transplantation (HSCT) [133-135], AML remains a major therapeutic challenge since relapse after CR is still the main obstacle and is difficult to manage due to patients' nonrandom heterogenous response to stereotypical treatment [128,136].

The nature of AML changes with patients age

AML can occur in people of all age. However, AML is primarily diagnosed among patients older than 60 years of age, with a median age of 68 years at diagnosis in the US [18]. Recent advances in AML biology that expanded our understanding of its complex genetic landscape and led to significant improvement in AML prognoses for younger patients. However, therapeutic strategies for AML patients have been nearly the same for more than 30 years [116,137], with almost no treatment options for patients older than 60 years of age [129,130]. Approximately 70% of patients older than 65 years of age die within

one year from diagnosis with current treatment[138].Additionaly, AML prognosis worsens as age increases due to increase in adverse cytogenetic abnormalities. Furthermore, response to treatments also worsens with age, with older patients respond less to treatments, with poorer clinical outcomes.

It is therefore unquestionable that the nature of AML changes with age, but despite this, little is known about the extent of these associations and how they vary with AML patient's age[22,129,139]. Thus, in the present study we seek to identify AML age-dependent and sex-related gene expression signatures by exploring the age-related gene expression patterns in AML.

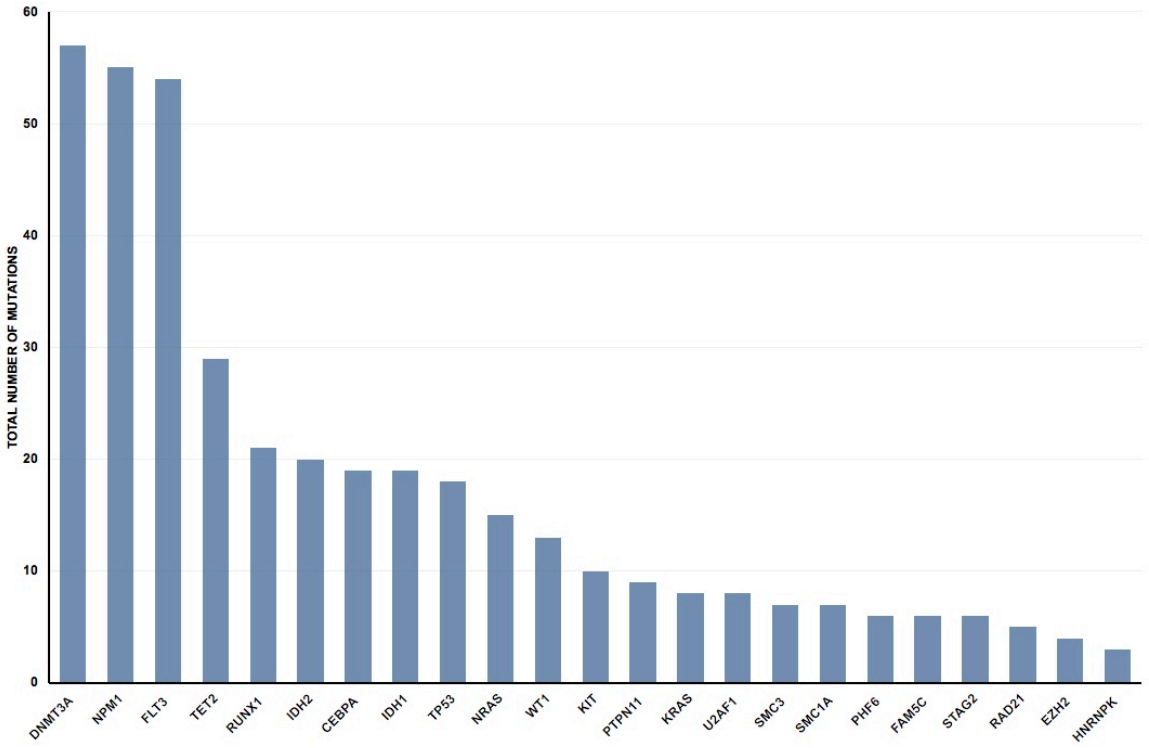
Conclusion

In this dissertation, we aimed to establish sex-linked and age-dependent biomarkers from genes with similar alteration in gene expression level and associated signaling pathway in AML. The approach utilized machine learning, which led to the development of a graphical user interface to facilitate model training and classification, (Chapter 2). Subsequently, meta-analyses of publicly available data were used to study the effects of age and sex in AML (Chapter 3). In particular, three analyses were performed to help us reach our aims **Analysis 1: "Gene expression meta-analysis and associated signaling pathways of AML disease state compared to healthy individuals"**, to identify differentially expressed (DE) genes in AML disease state, followed by gene enrichment analysis on the identified DE genes to find singling pathway associated with AML. **Analysis 2a: "Sex-relevance differential gene expression meta-analysis and associated signaling pathways in AML"**, to explore the relevance of patients' sex on gene expression and to identify sex-linked genes and associated signaling pathways in AML. **Analysis 2b: "Age-dependent gene expression meta-analysis and associated signaling pathways in AML",** to identify common set of age-dependent genes and associated signaling pathways and to explore age-dependent trends in AML gene expression. Finally, using our results and combined with a machine learning model (KNN model), we were able to classify AML patients compared to healthy individuals with > 90% achieved accuracy. Overall our findings provide a new reanalysis of public datasets, that enabled the identification of potential new gene sets relevant to AML that can potentially be used in future experiments and possible stratified disease diagnostics.

APPENDIX

APPENDIX

Figure 1. Genes frequently mutated in AML according to TCGA



The Cancer Genome Atlas Research Network (TCGA) analyzed the genomes of 200 de-novo AML patients. Analysis revealed a total of 23 genes (shown) that are frequently and significantly mutated in AML with 237 genes (not shown) were mutated in 2 or more samples in de novo AML. (data from reference 124)

Table 1. Classification of AML according to the WHO acute myeloid leukemia and related neoplasms classification system.

1: AML with recurrent genetic abnormalities
AML with t(8;21)(q22;q22.1);RUNX1-RUNX1T1
AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22);CBFB-MYH11
APL with PML-RARA
AML with t(9;11)(p21.3;q23.3);MLLT3-KMT2A
AML with t(6;9)(p23;q34.1);DEK-NUP214
AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2, MECOM
AML (megakaryoblastic) with t(1;22)(p13.3;q13.3);RBM15-MKL1
Provisional entity: AML with BCR-ABL1
AML with mutated NPM1
AML with biallelic mutations of CEBPA
Provisional entity: AML with mutated RUNX1
2: AML with myelodysplasia-related changes
3: Therapy-related myeloid neoplasms
4: AML, NOS
AML with minimal differentiation
AML without maturation
AML with maturation
Acute myelomonocytic leukemia
Acute monoblastic/monocytic leukemia
Pure erythroid leukemia
Acute megakaryoblastic leukemia

Table 2. Cytogenetic abnormalities sufficient for the diagnosis of AML with myelodysplasia-related changes (AML-MRC).

Complex karyotype (3 or more abnormalities)
None included in AML with recurrent genetic abnormalities
Unbalanced abnormalities
-7/del(7q)
del(5q)/t(5q)
i(17q)/t(17p)
-13/del(13q)
del(11q)
del(12p)/t(12p)
idic(X)(q13)
Balanced abnormalities
t(11;16)(q23;p13.3)
t(3;21)(q26.2;q21.2)
t(1;3)(p36.3;q21.1)
t(2;11)(p21;q23)
t(5;12)(q32;p13.2)
t(5;7)(q32;q11.2)
t(5;17)(q32;p13.2)
t(5;10)(q32;q21.2)
t(3;5)(q25.3;q35.1)

BIBLIOGRAPHY

BIBLIOGRAPHY

1 Sudhakar, A. History of Cancer, Ancient and Modern Treatment Methods. *J Cancer Sci Ther* **1**, 1-4, doi:10.4172/1948-5956.100000e2 (2009).

2 Piller, G. Leukaemia - a brief historical review from ancient times to 1950. *Br J Haematol* **112**, 282-292 (2001).

3 Gest, H. The remarkable vision of Robert Hooke (1635-1703): first observer of the microbial world. *Perspect Biol Med* **48**, 266-272, doi:10.1353/pbm.2005.0053 (2005).

4 Lane, N. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philos Trans R Soc Lond B Biol Sci* **370**, doi:10.1098/rstb.2014.0344 (2015).

5 Bennett, J. H. Case of hypertrophy of the spleen and liver, in which death took place from suppuration of the blood. *Edinburgh Med Sug J* **64**, 413-423 (1845).

6 Fuller, H. Particulars of a case in which enormous enlargement of the spleen and liver, together with dilation of all the blood vessels of the body were found co-incident with a peculiarly altered condition of the blood. *Lancet* **2**, 43-44 (1846).

7 Velpeau, A. Sur la resorption du pusaet sur l'alteration du sang dans les maladies clinique de persection nenemant. Premier observation. *Rev Med* **2**, 216 (1827).

8 Friedreich, N. Ein neuer Fall von Leukämie. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin* **12**, 37-58, doi:10.1007/bf01938747 (1857).

9 Ehrlich, P. Beiträge zur Kenntniss der Anilinfärbungen und ihrer Verwendung in der mikroskopischen Technik. *Archiv für mikroskopische Anatomie* **13**, 263-277, doi:10.1007/bf02933937 (1877).

10 Gowers, W. R. Splenic leucocythaemia. *A system of medicine* **5**, 216-305 (1879).

11 CABOT, R. C. Acute Leukemia. *The Boston Medical and Surgical Journal* **131**, 507-511, doi:10.1056/nejm189411221312103 (1894).

12 Buchanan, R. J. M. Flagellation of lymphocytes. *Brit Med J* **1909**, 306-306 (1909).

13 Farmer, P. *et al.* Expansion of cancer care and control in countries of low and middle income: a call to action. *Lancet* **376**, 1186-1193, doi:10.1016/S0140-6736(10)61152-X (2010).

14 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7-30, doi:10.3322/caac.21442 (2018).

15 Yamamoto, J. F. & Goodman, M. T. Patterns of leukemia incidence in the United States by subtype and demographic characteristics, 1997-2002. *Cancer Causes Control* **19**, 379-390, doi:10.1007/s10552-007-9097-2 (2008).

16 De Kouchkovsky, I. & Abdul-Hay, M. 'Acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood Cancer J* **6**, e441, doi:10.1038/bcj.2016.50 (2016).

17 Deschler, B. & Lubbert, M. Acute myeloid leukemia: epidemiology and etiology. *Cancer* **107**, 2099-2107, doi:10.1002/cncr.22233 (2006).

18 Institute, N. C. SEER Cancer Stat Facts: Acute Myeloid Leukemia (Percent of New Cases by Age Group). [https://seer.cancer.gov/statfacts/html/amyl.html]. ((accessed 11.30.18), 2011-2015).

19 Estey, E. & Dohner, H. Acute myeloid leukaemia. *Lancet* **368**, 1894-1907, doi:10.1016/S0140-6736(06)69780-8 (2006).

20 Kumar, C. C. Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes Cancer* **2**, 95-107, doi:10.1177/1947601911408076 (2011).

21 Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391-2405, doi:10.1182/blood-2016-03-643544 (2016).

22 Dohner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N Engl J Med* **373**, 1136-1152, doi:10.1056/NEJMra1406184 (2015).

23 Peterson, L. F. & Zhang, D. E. The 8;21 translocation in leukemogenesis. *Oncogene* **23**, 4255-4262, doi:10.1038/sj.onc.1207727 (2004).

24 Rowley, J. D. Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. *Ann Genet* **16**, 109-112 (1973).

25 Miyoshi, H. *et al.* The t(8;21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript. *EMBO J* **12**, 2715-2721 (1993).

26 Speck, N. A. & Gilliland, D. G. Core-binding factors in haematopoiesis and leukaemia. *Nature Reviews Cancer* **2**, 502-513, doi:10.1038/nrc840 (2002).

27 Shurtleff, S. A. *et al.* Heterogeneity in CBF beta/MYH11 fusion messages encoded by the inv(16)(p13q22) and the t(16;16)(p13;q22) in acute myelogenous leukemia. *Blood* **85**, 3695-3703 (1995).

28 Le Beau, M. M. *et al.* Association of an inversion of chromosome 16 with abnormal marrow eosinophils in acute myelomonocytic leukemia. A unique cytogenetic-clinicopathological association. *N Engl J Med* **309**, 630-636, doi:10.1056/NEJM198309153091103 (1983).

29 Liu, P. *et al.* Identification of yeast artificial chromosomes containing the inversion 16 p-arm breakpoint associated with acute myelomonocytic leukemia. *Blood* **82**, 716-721 (1993).

30 Liu, P. *et al.* Fusion between transcription factor CBF beta/PEBP2 beta and a myosin heavy chain in acute myeloid leukemia. *Science* **261**, 1041-1044 (1993).

31 Sinha, C., Cunningham, L. C. & Liu, P. P. Core Binding Factor Acute Myeloid Leukemia: New Prognostic Categories and Therapeutic Opportunities. *Semin Hematol* **52**, 215-222, doi:10.1053/j.seminhematol.2015.04.002 (2015).

32 Sood, R., Kamikubo, Y. & Liu, P. Role of RUNX1 in hematological malignancies. *Blood* **129**, 2070-2082, doi:10.1182/blood-2016-10-687830 (2017).

33 Estey, E. H. Acute myeloid leukemia: 2012 update on diagnosis, risk stratification, and management. *Am J Hematol* **87**, 90-99, doi:10.1002/ajh.22246 (2012).

34 Adams, J. & Nassiri, M. Acute Promyelocytic Leukemia: A Review and Discussion of Variant Translocations. *Arch Pathol Lab Med* **139**, 1308-1313, doi:10.5858/arpa.2013-0345-RS (2015).

35 Stone, R. M. & Mayer, R. J. The unique aspects of acute promyelocytic leukemia. *J Clin Oncol* **8**, 1913-1921, doi:10.1200/JCO.1990.8.11.1913 (1990).

36 Hillestad, L. K. Acute promyelocytic leukemia. *Acta Med Scand* **159**, 189-194 (1957).

37 Rowley, J. D., Golomb, H. M. & Dougherty, C. 15/17 translocation, a consistent chromosomal change in acute promyelocytic leukaemia. *Lancet* **1**, 549-550 (1977).

38 Grignani, F. *et al.* The acute promyelocytic leukemia-specific PML-RAR alpha fusion protein inhibits differentiation and promotes survival of myeloid precursor cells. *Cell* **74**, 423-431 (1993).

39 Lee, G. Y. *et al.* Acute promyelocytic leukemia with PML-RARA fusion on i(17q) and therapy-related acute myeloid leukemia. *Cancer Genet Cytogenet* **159**, 129-136, doi:10.1016/j.cancergencyto.2004.09.019 (2005).

40 Giguere, V., Ong, E. S., Segui, P. & Evans, R. M. Identification of a receptor for the morphogen retinoic acid. *Nature* **330**, 624-629, doi:10.1038/330624a0 (1987).

41 Petkovich, M., Brand, N. J., Krust, A. & Chambon, P. A human retinoic acid receptor which belongs to the family of nuclear receptors. *Nature* **330**, 444-450, doi:10.1038/330444a0 (1987).

42 de The, H. *et al.* The PML-RAR alpha fusion mRNA generated by the t(15;17) translocation in acute promyelocytic leukemia encodes a functionally altered RAR. *Cell* **66**, 675-684 (1991).

43 Kakizuka, A. *et al.* Chromosomal translocation t(15;17) in human acute promyelocytic leukemia fuses RAR alpha with a novel putative transcription factor, PML. *Cell* **66**, 663-674 (1991).

44 Swansbury, G. J., Slater, R., Bain, B. J., Moorman, A. V. & Secker-Walker, L. M. Hematological malignancies with t(9;11)(p21-22;q23)--a laboratory and clinical study of 125 cases. European 11q23 Workshop participants. *Leukemia* **12**, 792-800 (1998).

45 Schoch, C. *et al.* AML with 11q23/MLL abnormalities as defined by the WHO classification: incidence, partner chromosomes, FAB subtype, age distribution, and prognostic impact in an unselected series of 1897 cytogenetically analyzed AML cases. *Blood* **102**, 2395-2402, doi:10.1182/blood-2003-02-0434 (2003).

46 Bower, M. *et al.* Prevalence and clinical correlations of MLL gene rearrangements in AML-M4/5. *Blood* **84**, 3776-3780 (1994).

47 Hilden, J. M. & Kersey, J. H. The MLL (11q23) and AF-4 (4q21) genes disrupted in t(4;11) acute leukemia: molecular and clinical studies. *Leuk Lymphoma* **14**, 189-195, doi:10.3109/10428199409049668 (1994).

48 Marschalek, R. Systematic Classification of Mixed-Lineage Leukemia Fusion Partners Predicts Additional Cancer Pathways. *Ann Lab Med* **36**, 85-100, doi:10.3343/alm.2016.36.2.85 (2016).

49 Hess, J. L. MLL: a histone methyltransferase disrupted in leukemia. *Trends Mol Med* **10**, 500-507, doi:10.1016/j.molmed.2004.08.005 (2004).

50 Schwartz, S., Jiji, R., Kerman, S., Meekins, J. & Cohen, M. M. Translocation (6;9)(p23;q34) in acute nonlymphocytic leukemia. *Cancer Genet Cytogenet* **10**, 133-138 (1983).

51 Rowley, J. D. & Potter, D. Chromosomal banding patterns in acute nonlymphocytic leukemia. *Blood* **47**, 705-721 (1976).

52 Chi, Y., Lindgren, V., Quigley, S. & Gaitonde, S. Acute myelogenous leukemia with t(6;9)(p23;q34) and marrow basophilia: an overview. *Arch Pathol Lab Med* **132**, 1835-1837, doi:10.1043/1543-2165-132.11.1835 (2008).

53 Suzukawa, K. *et al.* Identification of a breakpoint cluster region 3' of the ribophorin I gene at 3q21 associated with the transcriptional activation of the EVI1 gene in acute myelogenous leukemias with inv(3)(q21q26). *Blood* **84**, 2681-2688 (1994).

54 Bitter, M. A., Neilly, M. E., Le Beau, M. M., Pearson, M. G. & Rowley, J. D. Rearrangements of chromosome 3 involving bands 3q21 and 3q26 are associated with normal or elevated platelet counts in acute nonlymphocytic leukemia. *Blood* **66**, 1362-1370 (1985).

55 Grigg, A. P., Gascoyne, R. D., Phillips, G. L. & Horsman, D. E. Clinical, haematological and cytogenetic features in 24 patients with structural rearrangements of the Q arm of chromosome 3. *Br J Haematol* **83**, 158-165 (1993).

56 Lugthart, S. *et al.* Clinical, molecular, and prognostic significance of WHO type inv(3)(q21q26.2)/t(3;3)(q21;q26.2) and various other 3q abnormalities in acute myeloid leukemia. *J Clin Oncol* **28**, 3890-3898, doi:10.1200/JCO.2010.29.2771 (2010).

57 Lugus, J. J. *et al.* GATA2 functions at multiple steps in hemangioblast development and differentiation. *Development* **134**, 393-405, doi:10.1242/dev.02731 (2007).

58 Buonamici, S., Chakraborty, S., Senyuk, V. & Nucifora, G. The role of EVI1 in normal and leukemic cells. *Blood Cells Mol Dis* **31**, 206-212 (2003).

59 Groschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381, doi:10.1016/j.cell.2014.02.019 (2014).

60 Yamazaki, H. *et al.* A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell* **25**, 415-427, doi:10.1016/j.ccr.2014.02.008 (2014).

61 Orazi, A. Histopathology in the diagnosis and classification of acute myeloid leukemia, myelodysplastic syndromes, and myelodysplastic/myeloproliferative diseases. *Pathobiology* **74**, 97-114, doi:10.1159/000101709 (2007).

62 Bennett, J. M. *et al.* Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br J Haematol* **33**, 451-458 (1976).

63 Lu, G., Altman, A. J. & Benn, P. A. Review of the cytogenetic changes in acute megakaryoblastic leukemia: one disease or several? *Cancer Genet Cytogenet* **67**, 81-89 (1993).

64 Lion, T. *et al.* The translocation t(1;22)(p13;q13) is a nonrandom marker specifically associated with acute megakaryocytic leukemia in young children. *Blood* **79**, 3325-3330 (1992).

65 Baruchel, A., Daniel, M. T., Schaison, G. & Berger, R. Nonrandom t(1;22)(p12-p13;q13) in acute megakaryocytic malignant proliferation. *Cancer Genet Cytogenet* **54**, 239-243 (1991).

66 Carroll, A. *et al.* The t(1;22) (p13;q13) is nonrandom and restricted to infants with acute megakaryoblastic leukemia: a Pediatric Oncology Group Study. *Blood* **78**, 748-752 (1991).

67 Ma, Z. *et al.* Fusion of two novel genes, RBM15 and MKL1, in the t(1;22)(p13;q13) of acute megakaryoblastic leukemia. *Nat Genet* **28**, 220-221, doi:10.1038/90054 (2001).

68 Novoyatleva, T. *et al.* Protein phosphatase 1 binds to the RNA recognition motif of several splicing factors and regulates alternative pre-mRNA processing. *Hum Mol Genet* **17**, 52-70, doi:10.1093/hmg/ddm284 (2008).

69 Mercher, T. *et al.* Recurrence of OTT-MAL fusion in t(1;22) of infant AML-M7. *Genes Chromosomes Cancer* **33**, 22-28 (2002).

70 Wiellette, E. L. *et al.* spen encodes an RNP motif protein that interacts with Hox pathways to repress the development of head-like sclerites in the Drosophila trunk. *Development* **126**, 5373-5385 (1999).

71 Konoplev, S. *et al.* Molecular characterization of de novo Philadelphia chromosome-positive acute myeloid leukemia. *Leuk Lymphoma* **54**, 138-144, doi:10.3109/10428194.2012.701739 (2013).

72 Keung, Y. K. *et al.* Philadelphia chromosome positive myelodysplastic syndrome and acute myeloid leukemia-retrospective study and review of literature. *Leuk Res* **28**, 579-586, doi:10.1016/j.leukres.2003.10.027 (2004).

73 Soupir, C. P. *et al.* Philadelphia chromosome-positive acute myeloid leukemia: a rare aggressive leukemia with clinicopathologic features distinct from chronic

myeloid leukemia in myeloid blast crisis. *Am J Clin Pathol* **127**, 642-650, doi:10.1309/B4NVER1AJJ84CTUU (2007).

74 Berger, R. Differences between blastic chronic myeloid leukemia and Ph-positive acute leukemia. *Leuk Lymphoma* **11 Suppl 1**, 235-237, doi:10.3109/10428199309047892 (1993).

75 Colicelli, J. ABL tyrosine kinases: evolution of function, regulation, and specificity. *Sci Signal* **3**, re6, doi:10.1126/scisignal.3139re6 (2010).

76 Diekmann, D. *et al.* Bcr encodes a GTPase-activating protein for p21rac. *Nature* **351**, 400-402, doi:10.1038/351400a0 (1991).

77 Maru, Y. & Witte, O. N. The BCR gene encodes a novel serine/threonine kinase activity within a single exon. *Cell* **67**, 459-468 (1991).

78 Laurent, E., Talpaz, M., Kantarjian, H. & Kurzrock, R. The BCR gene and philadelphia chromosome-positive leukemogenesis. *Cancer Res* **61**, 2343-2355 (2001).

79 Neuendorff, N. R., Burmeister, T., Dorken, B. & Westermann, J. BCR-ABL-positive acute myeloid leukemia: a new entity? Analysis of clinical and molecular features. *Ann Hematol* **95**, 1211-1221, doi:10.1007/s00277-016-2721-z (2016).

80 Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med* **352**, 254-266, doi:10.1056/NEJMoa041974 (2005).

81 Heath, E. M. *et al.* Biological and clinical consequences of NPM1 mutations in AML. *Leukemia* **31**, 798-807, doi:10.1038/leu.2017.30 (2017).

82 Federici, L. & Falini, B. Nucleophosmin mutations in acute myeloid leukemia: a tale of protein unfolding and mislocalization. *Protein Sci* **22**, 545-556, doi:10.1002/pro.2240 (2013).

83 Borer, R. A., Lehner, C. F., Eppenberger, H. M. & Nigg, E. A. Major nucleolar proteins shuttle between nucleus and cytoplasm. *Cell* **56**, 379-390 (1989).

84 Colombo, E., Alcalay, M. & Pelicci, P. G. Nucleophosmin and its complex network: a possible therapeutic target in hematological diseases. *Oncogene* **30**, 2595-2609, doi:10.1038/onc.2010.646 (2011).

85 Lindstrom, M. S. NPM1/B23: A Multifunctional Chaperone in Ribosome Biogenesis and Chromatin Remodeling. *Biochem Res Int* **2011**, 195209, doi:10.1155/2011/195209 (2011).

86 Okuwaki, M., Iwamatsu, A., Tsujimoto, M. & Nagata, K. Identification of nucleophosmin/B23, an acidic nucleolar protein, as a stimulatory factor for in vitro replication of adenovirus DNA complexed with viral basic core proteins. *J Mol Biol* **311**, 41-55, doi:10.1006/jmbi.2001.4812 (2001).

87 Dumbar, T. S., Gentry, G. A. & Olson, M. O. Interaction of nucleolar phosphoprotein B23 with nucleic acids. *Biochemistry* **28**, 9495-9501 (1989).

88 Bertwistle, D., Sugimoto, M. & Sherr, C. J. Physical and functional interactions of the Arf tumor suppressor protein with nucleophosmin/B23. *Mol Cell Biol* **24**, 985-996 (2004).

89 Colombo, E., Marine, J. C., Danovi, D., Falini, B. & Pelicci, P. G. Nucleophosmin regulates the stability and transcriptional activity of p53. *Nat Cell Biol* **4**, 529-533, doi:10.1038/ncb814 (2002).

90 Dohner, K. *et al.* 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, doi:10.1182/blood-2005-05-2164 (2005).

91 Pabst, T. *et al.* Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia. *Nat Genet* **27**, 263-270, doi:10.1038/85820 (2001).

92 Green, C. L. *et al.* Prognostic significance of CEBPA mutations in a large cohort of younger adult patients with acute myeloid leukemia: impact of double CEBPA mutations and the interaction with FLT3 and NPM1 mutations. *J Clin Oncol* **28**, 2739-2747, doi:10.1200/JCO.2009.26.2501 (2010).

93 Mrozek, K., Heinonen, K. & Bloomfield, C. D. Clinical importance of cytogenetics in acute myeloid leukaemia. *Best Pract Res Clin Haematol* **14**, 19-47, doi:10.1053/beha.2000.0114 (2001).

94 Calkhoven, C. F., Muller, C. & Leutz, A. Translational control of C/EBPalpha and C/EBPbeta isoform expression. *Genes Dev* **14**, 1920-1932 (2000).

95 Nerlov, C. C/EBPalpha mutations in acute myeloid leukaemias. *Nat Rev Cancer* **4**, 394-400, doi:10.1038/nrc1363 (2004).

96 Tenen, D. G., Hromas, R., Licht, J. D. & Zhang, D. E. Transcription factors, normal myeloid development, and leukemia. *Blood* **90**, 489-519 (1997).

97 Dufour, A. *et al.* Acute myeloid leukemia with biallelic CEBPA gene mutations and normal karyotype represents a distinct genetic entity associated with a

favorable clinical outcome. *J Clin Oncol* **28**, 570-577, doi:10.1200/JCO.2008.21.6010 (2010).

98 Pabst, T. & Mueller, B. U. Transcriptional dysregulation during myeloid transformation in AML. *Oncogene* **26**, 6829-6837, doi:10.1038/sj.onc.1210765 (2007).

99 Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088-3091, doi:10.1182/blood-2008-09-179895 (2009).

100 Pabst, T., Eyholzer, M., Fos, J. & Mueller, B. U. Heterogeneity within AML with CEBPA mutations; only CEBPA double mutations, but not single CEBPA mutations are associated with favourable prognosis. *Br J Cancer* **100**, 1343-1346, doi:10.1038/sj.bjc.6604977 (2009).

101 Ichikawa, M. *et al.* AML-1 is required for megakaryocytic maturation and lymphocytic differentiation, but not for maintenance of hematopoietic stem cells in adult hematopoiesis. *Nat Med* **10**, 299-304, doi:10.1038/nm997 (2004).

102 Sakurai, M. *et al.* Impaired hematopoietic differentiation of RUNX1-mutated induced pluripotent stem cells derived from FPD/AML patients. *Leukemia* **28**, 2344-2354, doi:10.1038/leu.2014.136 (2014).

103 Gaidzik, V. I. *et al.* RUNX1 mutations in acute myeloid leukemia: results from a comprehensive genetic and clinical analysis from the AML study group. *J Clin Oncol* **29**, 1364-1372, doi:10.1200/JCO.2010.30.7926 (2011).

104 Lindsley, R. C. *et al.* Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood* **125**, 1367-1376, doi:10.1182/blood-2014-11-610543 (2015).

105 Gaidzik, V. I. *et al.* RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. *Leukemia* **30**, 2160-2168, doi:10.1038/leu.2016.126 (2016).

106 Yanada, M. *et al.* Long-term outcomes for unselected patients with acute myeloid leukemia categorized according to the World Health Organization classification: a single-center experience. *Eur J Haematol* **74**, 418-423, doi:10.1111/j.1600-0609.2004.00397.x (2005).

107 Falini, B. *et al.* Multilineage dysplasia has no impact on biologic, clinicopathologic, and prognostic features of AML with mutated nucleophosmin (NPM1). *Blood* **115**, 3776-3786, doi:10.1182/blood-2009-08-240457 (2010).

108 Diaz-Beya, M. *et al.* The prognostic value of multilineage dysplasia in de novo acute myeloid leukemia patients with intermediate-risk cytogenetics is dependent on NPM1 mutational status. *Blood* **116**, 6147-6148, doi:10.1182/blood-2010-09-307314 (2010).

109 Bacher, U. *et al.* Multilineage dysplasia does not influence prognosis in CEBPA-mutated AML, supporting the WHO proposal to classify these patients as a unique entity. *Blood* **119**, 4719-4722, doi:10.1182/blood-2011-12-395574 (2012).

110 Rozman, M. *et al.* Multilineage dysplasia is associated with a poorer prognosis in patients with de novo acute myeloid leukemia with intermediate-risk cytogenetics and wild-type NPM1. *Ann Hematol* **93**, 1695-1703, doi:10.1007/s00277-014-2100-6 (2014).

111 Haferlach, C. *et al.* AML with mutated NPM1 carrying a normal or aberrant karyotype show overlapping biologic, pathologic, immunophenotypic, and prognostic features. *Blood* **114**, 3024-3032, doi:10.1182/blood-2009-01-197871 (2009).

112 Singh, Z. N. *et al.* Therapy-related myelodysplastic syndrome: morphologic subclassification may not be clinically relevant. *Am J Clin Pathol* **127**, 197-205, doi:10.1309/NQ3PMV4U8YV39JWJ (2007).

113 Arber, D. A. *et al.* Acute myeloid leukaemia and related precursor neoplasms. *WHO classification of tumours of haematopoietic and lymphoid tissues* **1**, 110-139 (2008).

114 McNerney, M. E., Godley, L. A. & Le Beau, M. M. Therapy-related myeloid neoplasms: when genetics and environment collide. *Nat Rev Cancer* **17**, 513-527, doi:10.1038/nrc.2017.60 (2017).

115 Walter, R. B. *et al.* Significance of FAB subclassification of "acute myeloid leukemia, NOS" in the 2008 WHO classification: analysis of 5848 newly diagnosed patients. *Blood* **121**, 2424-2431, doi:10.1182/blood-2012-10-462440 (2013).

116 Dohner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453-474, doi:10.1182/blood-2009-07-235358 (2010).

117 Grimwade, D. & Hills, R. K. Independent prognostic factors for AML outcome. *Hematology Am Soc Hematol Educ Program*, 385-395, doi:10.1182/asheducation-2009.1.385 (2009).

118 Dohner, H. Implication of the molecular characterization of acute myeloid leukemia. *Hematology Am Soc Hematol Educ Program*, 412-419, doi:10.1182/asheducation-2007.1.412 (2007).

119 Bullinger, L. *et al.* Identification of acquired copy number alterations and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis. *Leukemia* **24**, 438-449, doi:10.1038/leu.2009.263 (2010).

120 Walter, M. J. *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci U S A* **106**, 12950-12955, doi:10.1073/pnas.0903091106 (2009).

121 Suela, J., Alvarez, S. & Cigudosa, J. C. DNA profiling by arrayCGH in acute myeloid leukemia and myelodysplastic syndromes. *Cytogenet Genome Res* **118**, 304-309, doi:10.1159/000108314 (2007).

122 Martelli, M. P., Sportoletti, P., Tiacci, E., Martelli, M. F. & Falini, B. Mutational landscape of AML with normal cytogenetics: biological and clinical implications. *Blood Rev* **27**, 13-22, doi:10.1016/j.blre.2012.11.001 (2013).

123 Zaidi, S. Z. *et al.* The challenge of risk stratification in acute myeloid leukemia with normal karyotype. *Hematol Oncol Stem Cell Ther* **1**, 141-158 (2008).

124 Cancer Genome Atlas Research, N. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).

125 Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278, doi:10.1016/j.cell.2012.06.023 (2012).

126 Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* **366**, 1090-1098, doi:10.1056/NEJMoa1106968 (2012).

127 Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209-2221, doi:10.1056/NEJMoa1516192 (2016).

128 Yang, X. & Wang, J. Precision therapy for acute myeloid leukemia. *J Hematol Oncol* **11**, 3, doi:10.1186/s13045-017-0543-7 (2018).

129 Ferrara, F. & Schiffer, C. A. Acute myeloid leukaemia in adults. *Lancet* **381**, 484-495, doi:10.1016/S0140-6736(12)61727-9 (2013).

130 Isidori, A. *et al.* Alternative novel therapies for the treatment of elderly acute myeloid leukemia patients. *Expert Rev Hematol* **6**, 767-784, doi:10.1586/17474086.2013.858018 (2013).

131 Nazha, A. & Ravandi, F. Acute myeloid leukemia in the elderly: do we know who should be treated and how? *Leukemia Lymphoma* **55**, 979-987, doi:10.3109/10428194.2013.828348 (2014).

132 Kantarjian, H. *et al.* Intensive chemotherapy does not benefit most older patients (age 70 years or older) with acute myeloid leukemia. *Blood* **116**, 4422-4429, doi:10.1182/blood-2010-03-276485 (2010).

133 Lowenberg, B. *et al.* Cytarabine dose for acute myeloid leukemia. *N Engl J Med* **364**, 1027-1036, doi:10.1056/NEJMoa1010222 (2011).

134 Lowenberg, B. Sense and nonsense of high-dose cytarabine for acute myeloid leukemia. *Blood* **121**, 26-28, doi:10.1182/blood-2012-07-444851 (2013).

135 Lowenberg, B. *et al.* High-dose daunorubicin in older patients with acute myeloid leukemia. *N Engl J Med* **361**, 1235-1248, doi:10.1056/NEJMoa0901409 (2009).

136 Estey, E. Acute Myeloid Leukemia - Many Diseases, Many Treatments. *N Engl J Med* **375**, 2094-2095, doi:10.1056/NEJMe1611424 (2016).

137 Reese, N. D. & Schiller, G. J. High-dose cytarabine (HD araC) in the treatment of leukemias: a review. *Curr Hematol Malig Rep* **8**, 141-148, doi:10.1007/s11899-013-0156-3 (2013).

138 Meyers, J., Yu, Y., Kaye, J. A. & Davis, K. L. Medicare fee-for-service enrollees with primary acute myeloid leukemia: an analysis of treatment patterns, survival, and healthcare resource utilization and costs. *Appl Health Econ Health Policy* **11**, 275-286, doi:10.1007/s40258-013-0032-2 (2013).

139 Appelbaum, F. R. *et al.* Age and acute myeloid leukemia. *Blood* **107**, 3481-3485, doi:10.1182/blood-2005-09-3724 (2006).

Chapter 2 –

ClassificaIO: machine learning for classification graphical user interface

Abstract

Machine learning methods are being used routinely by scientists in many research areas, typically requiring significant statsistical and programing knowledge. Here we present ClassificaIO, an open-source Python graphical user interface for machine learning classification for the scikit-learn Python library. ClassificaIO provides an interactive way to train, validate, and test data on a range of classification algorithms. ClassificaIO's core aim is to provide a point and click graphical user interface to enable fast comparisons within and across classifiers, and facilitates uploading and exporting of trained models, and both validation and testing data results to maximize machine learning utility in a shorter time instead of to writing a script for each task. ClassificaIO can also be an educational tool that can enable biomedical and other researchers with minimal machine learning background to apply machine learning algorithms to their research in an interactive point-and-click way. For this thesis, the primary motivation for creating and utilizing ClassificaIO was to address missing annotations in AML publicly available microarray data. Our aim was to train multiple machine learning classifiers and use them to predict such missing annotations, including sex and sample source (tissue type), from curated publicly available AML gene expression data that could be used in a meta-analysis of a large cohort of studies on AML (Chapter 3)

The ClassificaIO package is available for download and installation through the Python Package Index (PyPI) (http://pypi.python.org/pypi/ClassificaIO) and it can be deployed using the "import" function in Python once the package is installed. The application is distributed under an MIT license and the source code is publicly available for download

(for Mac OS X, Linux and Microsoft Windows) through PyPI and GitHub (http://github.com/gmiaslab/ClassificaIO, and https://doi.org/10.5281/zenodo.1320465).

A version of this chapter and results has been submitted for publication.

Introduction

Recent advances in high-throughput technologies, especially in genomics, have led to an explosion of large-scale structured data (e.g. RNA-sequencing and microarray data) [1]. Machine learning methods (classification, regression, clustering, etc.) are routinely used in mining such big data to extract biological insights in a range of areas within genetics and genomics [2]. For example, using unsupervised machine learning classification methods to predict the sex of gene expression microarrays donor samples [3], using genome sequencing data to train machine learning models to identify transcription start sites [4], splice sites [5], transcriptional promoters and enhancers regions [6]. Recent examples of using machine learning classification methods include their use to detect neurofibromin 1 tumor suppressor gene inactivation in glioblastoma [7], and to identify reliable gene markers for drug sensitivity in acute myeloid leukemia [8]. Many advanced machine learning algorithms have been developed in the recent years. Scikit-learn [9] is one of the most popular machine learning libraries in Python with a plethora of thoroughly tested and well-maintained machine learning algorithms. However, these algorithms are primarily aimed at users with computational and statistical backgrounds, which may discourage many biologists, biomedical scientists or beginning students (who may have minimal machine learning background but still want to explore its application in their research) from using machine learning. Ching et al (2018) recently highlighted the role of deep learning (a class of machine learning algorithms) currently plays in biology, and how such algorithms present new opportunities and obstacles for a data-rich field such as biology [10].

Several open source machine learning applications, such as KNIME [11] and Weka [12] written in Java and Orange [13] written in Python, have been developed with graphical user interfaces. The dataflow process for most of these applications is generally graphically constructed by the user, in the form of placing and connecting widgets by drag-and-drop. Such graphical workflow and representation of data input, processing and output is visually appealing, but can be computationally demanding (memory, storage, processing, etc.) and limiting in algorithm comparison, since each machine learning algorithm can have many different parameters. These tools are very mature with numerous algorithms, and well documented. However, they can be intimidating for machine learning beginners and students that want to preform simple tasks such as data classification. Also, scikit-learn has comprehensive documentation [14], and many online resources, including though Kaggle [15] and Stack Overflow [16], and a large online user base, which make scikit-learn a very popular package for machine learning beginners learning using Python.

Here, we present ClassificaIO, an open-source Python graphical user interface (GUI) for supervised machine learning classification for the scikit-learn library. To our knowledge, no standalone GUI exists for the scikit-learn machine learning library. The core aim of ClassificaIO is to provide our research group with point and click graphical user interface to enable fast comparisons within and across different machine learning classifiers to predict/classify missing annotations, including sex and sample source, from our curated microarray data (for more details, see Chapter 3 section, **"Classification of missing metadata annotation"**). 805 and 737 arrays were missing for sex and sample source annotations from our curated data respectively. Since these arrays correspond to AML

patients and healthy individuals, the prediction of the missing annotations for these arrays was essential to our study for the statistical power and sample size.

ClassificaIO can also serve as a software tool for teaching and educational tool that is visually minimalistic and computationally light interactive interface, that can give access to a range of state-of-the-art classification algorithms to machine learning beginners with some basic knowledge of Python and using a terminal, and with broad background in machine learning, allowing them to use machine learning and apply it to their research. What distinguishes ClassificaIO from other similar applications is:

1. Cross-platform implementation for Mac OS X, Linux, and Windows operating systems

2. Interactive point-and-click GUI to 25 supervised classification algorithms in scikit-learn

3. Accessible clickable links, to scikit-learn's well-written online documentation for each implemented classification algorithms

4. Simple upload of all data files with dedicated buttons; with robust CSV reader, and a displayed history-log to track uploaded files, files names and directories

5. Fast comparisons within and across classifiers to train, validate, and test data

6. Upload and export of ClassificaIO trained models (for future of a trained model without the need to retrain), and export of both validated, and tested data results

7. Small application footprint in terms of disk space usage (<2 MB)

ClassificaIO implementation

ClassificaIO has been developed using the standard Python interface Tkinter module to the Tk GUI toolkit [17], for Mac OS X (using High Sierra ≥ 10.13), Linux (using Ubuntu 18.04-64 bit), and Windows (using Windows 10 64-bit). It uses external packages including: Tkinter, Pillow, Pandas [18], NumPy [19], scikit-learn and SciPy [20]. To avoid any system errors, crashes, and crude fonts, we recommend not to install ClassificaIO using integrated environment package installers – instead, native installation of ClassificaIO and dependencies (using pip for Mac and Windows, and pip3 and apt-get for Linux) is encouraged. Once installed, ClassificaIO can be deployed using the 'import' function. A ClassificaIO installation instruction and step by step working examples is distributed with ClassificaIO GUI and can be accessed directly through the 'HELP' button at the upper left of the GUI, that points the user's default browser to ClassificaIO's online user manual on GitHub. Some basic knowledge of Python and accessing it through a terminal are required for installation and running the software. Link to all supplementary files and additional ClassificaIO software information is provided in Table 3.

ClassificaIO backend

ClassificaIO implements 25 scikit-learn classification algorithms for supervised machine learning. A list of all these algorithms, their corresponding scikit-learn functions, and immutable (unchangeable) parameters with their default values are presented in Table 4, and ClassificaIO's workflow is outlined in Figure 2. Once training and testing data are uploaded to the front-end as described below, a classifier selection is made and submitted, ClassificaIO's backend calls the scikit-learn selected classifier, including any

values from manually set parameters to create the model. Otherwise, the default parameters values are used instead. For example, for "LogisticRegression", the model is defined in the scikit-learn library as a class, in terms of Python code used in the backend, the details are outlined in the scikit-learn documentation:

sklearn.linear_model.LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='liblinear', max_iter=100, multi_class='ovr', verbose=0, warm_start=False, n_jobs=)

The inputs to the class, within the parentheses, such as "penalty", "dual", "tol", etc., correspond to the model parameters, followed by an equal sign assigning the default values for these parameters. Rather than typing the values, the ClassificaIO GUI displays these parameters with input fields and radio buttons, for each classifier, initially populated by the default values. More information is available for all the parameters in the GUI, through a link for each classifier in the interface named "Learn More". The link directs the default browser to the scikit-learn online documentation of the selected classifier, and connects to the underlying backend documentation, and online parameter descriptions. The details and code complexity of the backend implementation are effectively hidden from the user, who can interact with the ClassificaIO GUI to set the relevant parameters, or leave them unchanged as default values. On the training data, ClassificaIO fits the estimator for classification using the scikit-learn 'fit' method, e.g. fit(x_train, y_train), to train (learn from the model), and uses the scikit-learn 'predict' method, e.g. predict(x_validation), to validate the model. Finally, ClassificaIO predicts new values using the scikit-learn 'predict' method again but on the testing data, e.g.

predict(testing_X), for implementing the model on new data that have not been used in model training.

ClassificaIO functionalities

ClassificaIO's GUI consists of three windows: 'Main', 'Use My Own Training Data', and 'Already Trained My Model'. Each window is actually implemented within the code as a class with several functions/methods that are dynamically connected to provide the GUI. ClassificaIO's Main window (Fig. 2) has two buttons: (i) the 'Use My Own Training Data' button, which when clicked allows the user to train and test classifiers using their own training and testing data, (ii) the 'Already Trained My Model' button, which when clicked allows the user to use their own already ClassificaIO trained model and testing data.

- **Data input**

For the 'Use My Own Training Data' window (Fig 4a) by clicking the corresponding buttons in the 'UPLOAD TRAINING DATA FILES' and 'UPLOAD TESTING DATA FILE' panels, a file selector directs the user to upload all required comma-separated values (CSV) data files ('Dependent and Target' or 'Dependent, Target and Features' and 'Testing Data') (Fig. 5). A history of all uploaded data files (file name and directory) is automatically saved in the 'CURRENT DATA UPLOAD' panel (Fig. 6). Briefly, the dependent data represent the data on which the model will depend on for learning, and the target data is the annotation, i.e. what is going to be predicted. The dependent data have attributes (also known as features) that take values (measurements/results) for each

contained object (i.e. each sample). Further details on data files formats and examples are provided in the Figures 7(a, b) and S2-S5.

- **Classifier selection**

After the data is uploaded, the user can select between all 25 different widely used classification algorithms (Table 4) (including logistic regression, perceptron, support vector machines, k-nearest neighbors, decision tree, random forest, neural network multi-layer perceptron, and more). The algorithms are integrated from the scikit-learn library, and allow the user to train and test models using their own uploaded data. Each classifier can be easily selected by clicking the corresponding classifier name in the 'CLASSIFER SELECTION' panel. Once classifier selection is completed, a brief description for the classifier with an underlined clickable link that reads "Learn more" right next to the classifier name (Fig. 8a) and the classifier parameters will populate. If "Learn more" is clicked, the link directs the default web browser to open scikit-learn's online well-written documentation that explains the specific classifier parameters, with explanation for each parameter and its use, and how to tune/optimize each parameter to get the best performance. ClassificaIO provides the user with an interactive point-and-click interface to set, modify, and test the influence of each parameter on their data (Fig. 8c). The user can switch between classifiers and parameters through point-and-click, which enables fast comparisons within and across classifier models.

- **Model training**

Both train-validate split and cross-validation methods (which are necessary to prevent/minimize overfitting) will populate with each classifier that can be used for data training (Fig. 8b). Training, validating and testing are all performed after pressing the submit button.

- **Results output**

After model training and testing is completed, the confusion matrix, classifier accuracy and error are displayed in the 'CONFUSION MATRIX, MODEL ACCURACY & ERROR' panel, bottom of (Fig. 9a). Model validation data results are displayed in the 'TRAINING RESULT: ID – ACTUAL – PREDICTION' panel (Fig. 9b), and testing data results are displayed in the 'TESTING RESULT: ID – PREDICTION' panel (Fig. 10b).

- **Model export**

By clicking on the 'Export Model' button, Figure 10 bottom left, the user can export trained models to save for future use without having to retrain. A previously exported ClassificaIO model can then be used for testing of new data in the 'Already Trained My Model' window (Fig. 4b) by clicking the 'Model file' button in the 'UPLOAD TRAINING MODEL FILE' panel (Fig 11a).

- **Results export**

Full results (trained models, both validated and tested data, and uploaded files names and directories) for both windows (Fig 4a&b), can be exported as CSV files for further analysis for publication, sharing, or later use (for more details on the exported trained model and data file formats, see S7 and S8).

Results: illustrative examples and data used

To illustrate the use of the interface and classification, we have used in this manuscript the following two examples.

- **Iris prediction using Iris dataset**

To demonstrate the interface and classification, we used the so-called Fisher/Anderson iris dataset [21,22]. This dataset is used widely as a prototype to illustrate classification algorithms, not only of biological data but in general machine learning implementations. The dataset consists of fifty samples each for three different species of iris flowers (Setosa, Versicolor and Viginica), with sepal length and width, and petal length and width provided as measurements. For more details on the iris data files format (**Fig 4a&b** and **S2-S5**).

- **Sex prediction using microarray gene expression data**

In this example, provided we used raw microarray gene expression data, from Gene Expression Omnibus (GEO) [23] to predict each sample donor's sex. This is often necessary

in metadata analyses, using publicly available gene expression datasets for reanalysis, as samples annotations on GEO may be missing information, including sample donor's sex. To illustrate the classification/sex prediction we used two datasets, GSE99039 [24] (training data) and GSE18781 [25] (testing data). In both GSE99039 and GSE18781 datasets, we used 121 and 25 samples respectively, for which RNA from peripheral blood mononuclear cells was assayed using Affymetrix Human Genome U133 Plus 2.0 Array (accession GPL570). The Y chromosome gene expression values were used in ClassificaIO as training and testing data to predict samples donor's sex. Using the 'Linear SVC' model with "k-fold cross validation" (10-fold), resulted into a model with 99% accuracy for sample donor's sex prediction (in the displayed example). For more details on the pre-processing of the raw gene expression data, files format, and Y chromosome probes ids, and final result see Ex2, **"Gene expression sex prediction using linear support vector classifier"** and Figures 10 &11.

Discussion

We have presented ClassificaIO, a GUI that implements the scikit-learn supervised machine learning classification algorithms. The scikit-learn package is one of the most popular in Python with well-written documentation, and many of its machine leaning algorithms are currently used for analyzing large and complex data sets in genomics. Our interface aims to provide an interactive machine learning research, teaching and educational tool to do machine learning analysis without the requirement of advanced computational and machine learning knowledge using scikit-learn. ClassificaIO is provided as an open source software, and its back-end classes and functions allow for

rapid development. We anticipate further development, aided by the scikit-learn library developer community to integrate additional classification algorithms, and extend ClassificaIO to include other machine leaning methods such as regression, clustering, and anomaly detection, to name but a few.

ClassificaIO: setup, dependencies, installation, instruction, and, step by step working examples

Summary

ClassificaIO is an open-source Python graphical user interface (GUI) for supervised machine learning classification for the scikit-learn module [9]. ClassificaIO aims to provide an easy-to-use interactive way to train, validate, and test data on a range of classification algorithms. The GUI enables fast comparisons within and across classifiers, and facilitates uploading and exporting of trained models, and both validated, and tested data results.

Dependencies

ClassificaIO is a Python package with the following external dependencies:

- Tkinter ≥ 8.6.7, Pillow ≥ 5.3.0, pandas ≥ 0.23.3, numpy == 1.15.3, scikit-learn ≥ 0.20.0, and scipy ≥ 1.1.0

Prerequisites

ClassificaIO requires Python version 3.6 or higher and can be used on Mac OS X High Sierra, Linux (tested on Ubuntu), and Windows 10 operating systems. To avoid any system errors, crashes, and crude fonts, we recommend to not install ClassificaIO using integrated environment package installers – i.e. native installation of ClassificaIO is highly encouraged using pip. In case you do not have pip installed, you must install it first.

Installation instructions

1. **Mac or Windows**

- To install the current release use pip in the terminal:

  $ pip install ClassificaIO

- Alternatively, you can install directly from GitHub using:

  $ pip install git+https://github.com/gmiaslab/ClassificaIO/

2. **Linux**

- First install the current release of tkinter and pip:

  $ sudo apt-get install python3-tk

  $ sudo apt-get install python3-pip

- To install the current ClassificaIO release use pip:

  $ pip3 install ClassificaIO

- Alternatively, you can install directly from GitHub using:

  $ pip install git+https://github.com/gmiaslab/ClassificaIO/

After installing ClassificaIO, please run it from the terminal using Python:

$ python3

>>> from ClassificaIO import ClassificaIO

>>> ClassificaIO.gui()

We note here the name is case sensitive (i.e. the 'IO' is capitalized). Once ClassificaIO's main window appears on your screen, you can click on 'Use My Own Training Data' button and start your supervised machine learning classification project.

Iris dataset prediction using a logistic regression classifier

- **Training data input**

You first need to select (either 'Dependent and Target' or 'Dependent, Target and Features') from the 'UPLOAD TRAINING DATA FILES' panel to upload training data files. For this example, we select the 'Dependent and Target' button.

To begin uploading data files, click the corresponding buttons in the 'UPLOAD TRAINING DATA FILES' panel: a file selector (Fig. 5) directs you to upload both, dependent or target data file. Once a file is uploaded to ClassificaIO, the file name and directory are automatically saved in the 'CURRENT DATA UPLOAD' panel (Fig. 6). Dependent data file (e.g. Fig. 7a) and target data file (e.g. Fig. 8b). This updatable log allows for tracking current data files in-use, and maintains a history of all files uploaded to the software.

- **Data format**

Data formats are shown in Figure 7a for dependent data and Figure 7b for target data. The dependent data represent the data on which the model will depend on for learning and the target data is the annotation, i.e. what is going to be predicted. In this example, the dependent data have 4 rows and 105 columns. For the dependent data, each row is an attribute (also known as feature) and each column is an object (also known as an observation or a sample). Thus, the header row enumerates the objects, and the header column names the attributes. The values in the file represent the measurement made for

each of the objects(columns) for each of the attributes (rows). For the target data, we have 105 rows and 2 columns (note: for the target data, the rows correspond to the objects and column is the class per object). The values in the "ids" column in the target data must much the Objects header row in the dependent data, the columns headers must match, otherwise an error will occur. Hence, the number of columns (i.e. objects) in the dependent data must also match the number of rows in the target data (i.e. each object has a unique "id" and must be assigned a target class for training). Finally, the "target" column in the target data must be numerically-valued.

- **Classifier selection**

Once you have uploaded all required training data files, you can select between 25 different machine learning classification algorithms in the 'CLASSIFER SELECTION' panel (Table 4). Here are all classification algorithms in order of appearance in the 'CLASSIFER SELECTION' panel. Immutable (unchangeable) parameters with their default values are also listed for each classifier in the parentheses:

Linear_model

LogisticRegression. (class_weight = None)

PassiveAggressiveClassifier. (class_weight = None, n_iter= None)

Perceptron. (class_weight = None)

RidgeClassifier. (class_weight = None)

Stochastic Gradient Descent (SGDClassifier).

Discriminant_analysis

LinearDiscriminantAnalysis. (shrinkage= None, priors = None)

QuadraticDiscriminantAnalysis. (store_covariances = None, priors = None)

Support vector machines (SVMs)

LinearSVC. (class_weight = None)

NuSVC. (class_weight = None)

SVC. (class_weight = None)

Neighbors

KNeighborsClassifier. (metric_params = None)

NearestCentroid.

RadiusNeighborsClassifier. (metric_params = None)\

Gaussian_process

GaussianProcessClassifier. (kernel = None)

Naive_bayes

BernoulliNB. (class_prior = None)

GaussianNB. (class_prior = None)

MultinomialNB. (class_prior = None)

Trees

DecisionTreeClassifier. (class_weight = None)

ExtraTreeClassifier . (min_impurity_split = None, class_weight = None)

Ensemble

AdaBoostClassifier. (base_estimator = None)

BaggingClassifier. (base_estimator = None)

ExtraTreesClassifier. (class_weight = None)

RandomForestClassifier. (class_weight = None)

Semi_supervised

LabelPropagation.

Neural_network

MLPClassifier.

The following will populate once you make a classifier selection:

Figure 8a: The classifier definition with a clickable underlined link “learn more” in blue, which, when clicked opens an external web-browser to the scikit-learn documentation for the selected classifier.

Figure 8b: Interactive way to select between train-validate split and cross-validation methods (radio buttons), which are necessary to prevent/minimize training model overfitting.

Figure 8c: Classifier parameters, to provide you with a point-and-click interface to set, modify, and test the influence of each parameter on your data

- **Model training, evaluation, validation and result output**

You can now click ‘submit’ to train your classifier using the uploaded training data files ‘Dependent and Target’ in this example, and evaluate your result. Or, alternatively you can upload testing data first, and then click ‘submit’ to train and test a classifier on your uploaded data at the same time. For this example, **first**: we train a selected classifier, ‘LogisticRegression’, using its default parameters, and default train-validate split method ‘Train Sample Size (%)’, and then, **second**: we upload testing data to test the trained model. After clicking ‘submit’, our selected classifier, ‘LogisticRegression’ for this

example, is trained using the loaded training data, 'Dependent and Target' for this example.

Notes: ClassificaIO always shuffles your training data before splitting to eliminate mini batch effects.

Internally, when 'Train Sample Size (%)' method is selected, ClassificaIO uses the scikit-learn *train_test_split* method, to allow for fast training data split into training and validation subsets. With this method the parameter set is *train_size*, which takes the train sample size set by you (e.g. Train Sample Size (%): set to 75% means *train_size* = 0.75 and *test_size* = 0.25). If the 'K-fold Cross-Validation' method is selected instead, ClassificaIO uses the scikit-learn *cross_val_predict* method where the training data is split into k-sets. The model is trained on k-1 of the folds followed by a validation step on the remaining part of the data. This will be repeated for each of the k-folds.

After training is completed, the confusion matrix, classifier accuracy and error are displayed in the 'CONFUSION MATRIX, MODEL ACCURACY & ERROR' panel (Fig. 9a). Model validation data results are displayed in the 'TRAINING RESULT: ID – ACTUAL – PREDICTION' panel (Fig. 6b) with each data point ID is the 1st value, actual target value is displayed 2nd, and predicted target value 3rd , where the predictions correspond to the iris flower species, with 0=setosa, 1=versicolor, and 2=virginica.

- **Testing data input and result output**

To test your trained model, first upload the testing data file by clicking the 'Testing Data' button in the 'UPLOAD TESTING DATA FILE' panel (Fig. 10a). Once clicked, a file

selector directs you to upload the testing data file, the file name is automatically saved in the 'CURRENT DATA UPLOAD' panel (outlined in the red box in the figure) to indicate that your file has been uploaded. The Testing Data file format is the same as for the dependent data file.

After clicking 'Submit', testing results are displayed in the 'TESTING RESULT: ID – PREDICTION' panel (Fig 7b) with each data point ID shown 1st, and the corresponding predicted target value displayed after it 2nd, separated by a hyphen.

- **Result export**

Now you are ready to export your trained model to preserve it for future use without having to retrain. Simply, click the 'Export Model' button (**Fig 6a**) and save your model. Your exported ClassificaIO model can then be used for future testing on new data in the 'Already Trained My Model' window in ClassificaIO, shown below.

- **ClassificaIO model input**

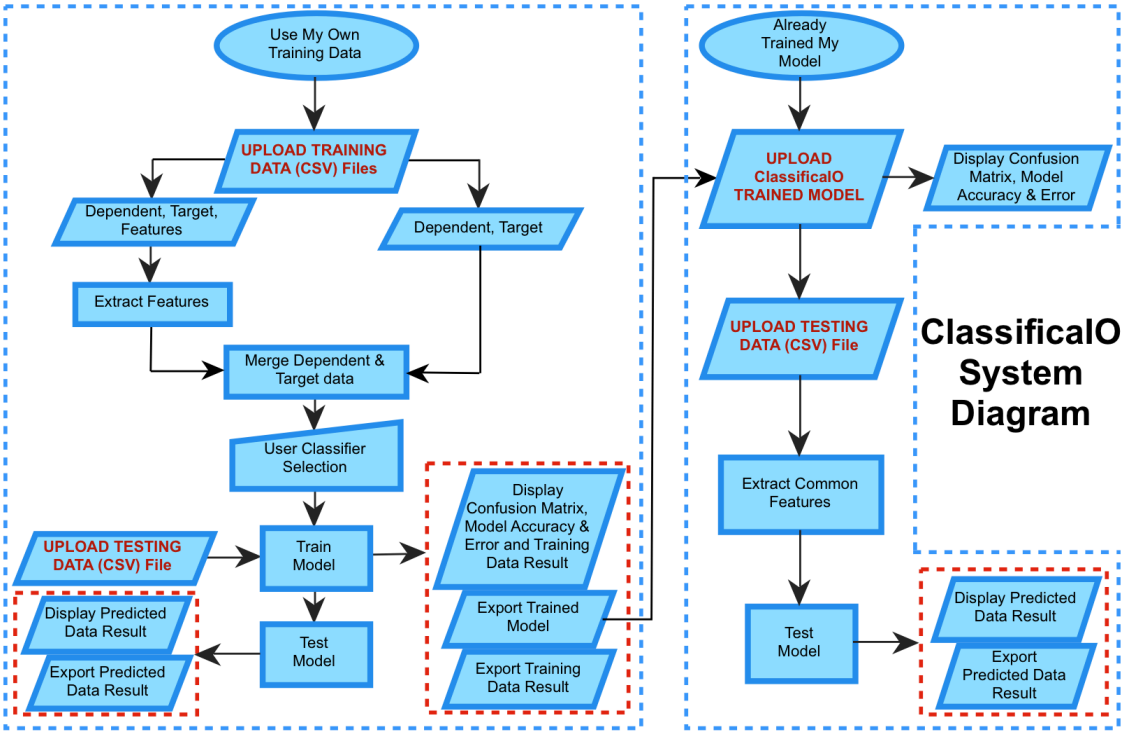
You will need to upload ClassificaIO model by clicking the 'Model File' button in the 'UPLOAD TRAINING MODEL FILE' panel (Fig. 11a). Once clicked, a file selector directs you to upload a ClassificaIO trained model. Also, you will need to upload a testing data file (the testing data file format is the same as explained above), by clicking the 'Testing Data' button in the "UPLOAD TESTING DATA FILE" panel (Fig. 11b). Once a ClassificaIO model and testing data files are uploaded, files names are automatically displayed in the 'CURRENT DATA UPLOAD' panel (Fig. 11c).

After clicking ‘submit’, the uploaded model preset parameters will populate (Fig. 11d) to show the classifier used to originally train the uploaded model. The confusion matrix, classifier accuracy and error of trained model are then displayed in the ‘CONFUSION MATRIX, MODEL ACCURACY & ERROR’ panel (Fig. 11e). Testing data results are displayed in the ‘Testing RESULT: ID – PREDICTION’ panel (Fig. 11f) with the data point ID shown 1st, followed by a hyphen and the predicted value displayed right after it.

APPENDIX

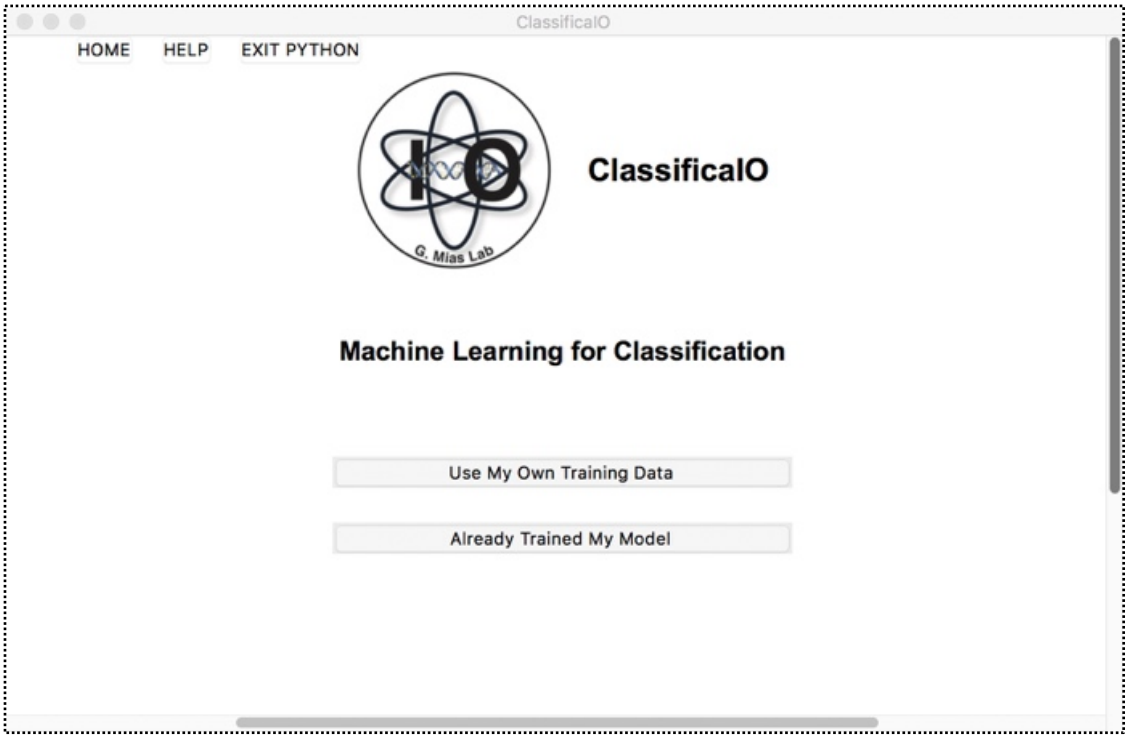
APPENDIX

Figure 2. Workflow summary of ClassificaIO.



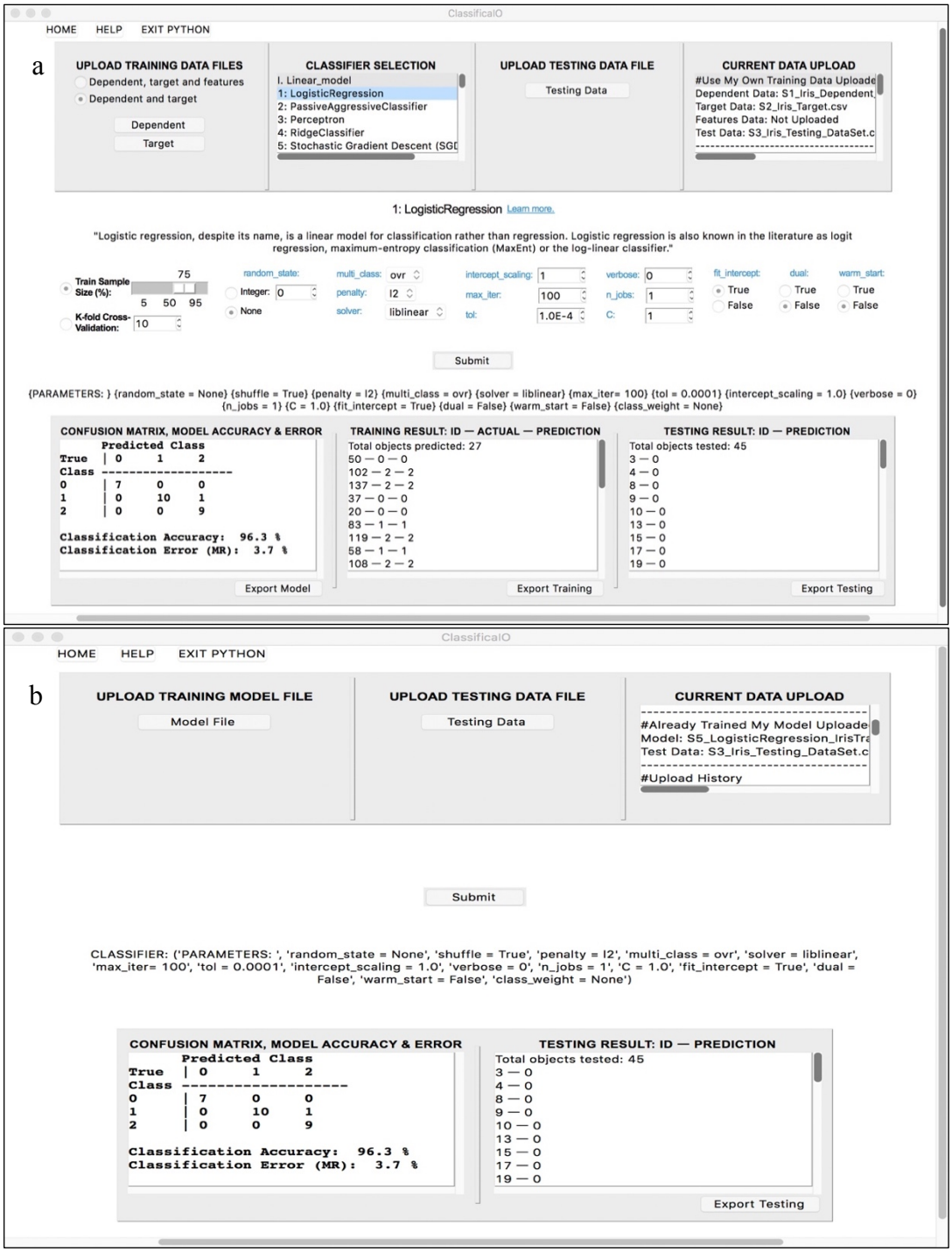
The diagram summarizes of the graphical user interface and backend functionality/workflow for ClassificaIO Use My Own Training Data window and Already Trained My Model window.

Figure 3. ClassificaIO main window.



ClassificaIO main window appears on the screen after typing 'ClassificaIO.gui()' in a terminal or a Python interpreter.

Figure 4. ClassificaIO user interface (Mac OS shown).

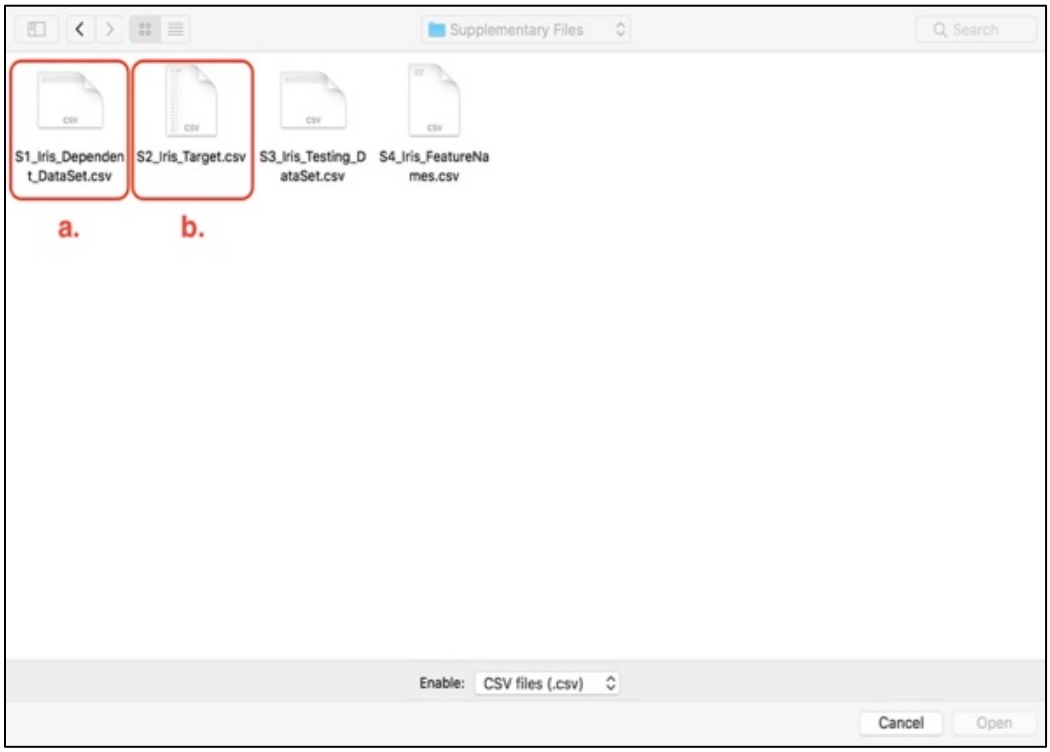


As described in ClassificaIO implementation section, **a.** an example Use My Own Training Data window with uploaded training and testing data files, selected logistic regression classifier, populated classifier parameters, and output classification results.

Figure 4. (cont'd)

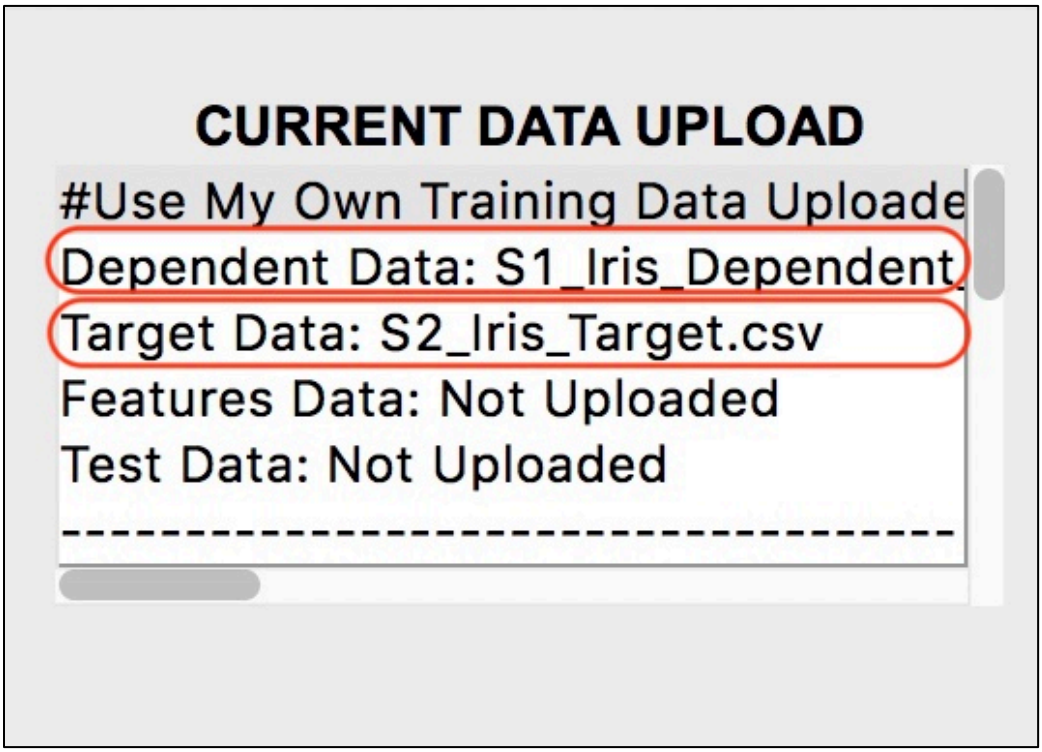
b. A corresponding Already Trained My Model window with uploaded ClassificaIO logistic regression trained model and testing data file, and output classification result.

Figure 5. Graphical control element dialog box.



a. Dependent data file selected for upload. **b.** Selected target data file to upload. N.B. each file selection has to be done one at a time.

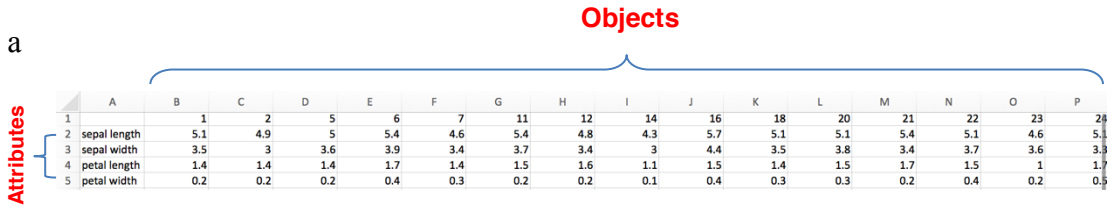
Figure 6. Current data upload panel.



Both dependent and target data file names shown (red boxes). Scroll down for uploaded data files directories.

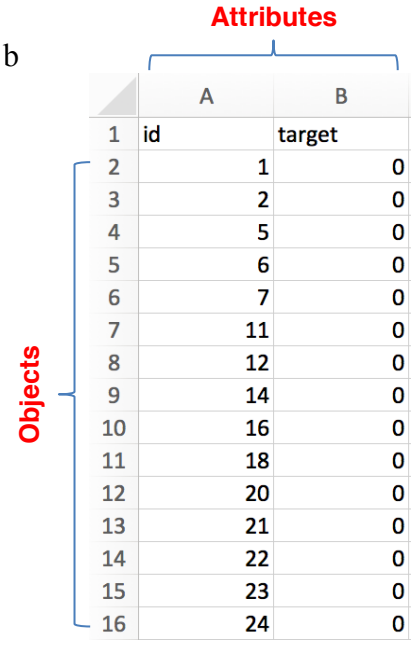
Figure 7. Gene expression sex prediction using linear support vector classifier.

a



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		1	2	5	6	7	11	12	14	16	18	20	21	22	23	24
2	sepal length	5.1	4.9	5	5.4	4.6	5.4	4.8	4.3	5.7	5.1	5.1	5.4	5.1	4.6	5.1
3	sepal width	3.5	3	3.6	3.9	3.4	3.7	3.4	3	4.4	3.5	3.8	3.4	3.7	3.6	3.3
4	petal length	1.4	1.4	1.4	1.7	1.4	1.5	1.6	1.1	1.5	1.4	1.5	1.7	1.5	1	1.7
5	petal width	0.2	0.2	0.2	0.4	0.3	0.2	0.2	0.1	0.4	0.3	0.3	0.2	0.4	0.2	0.5

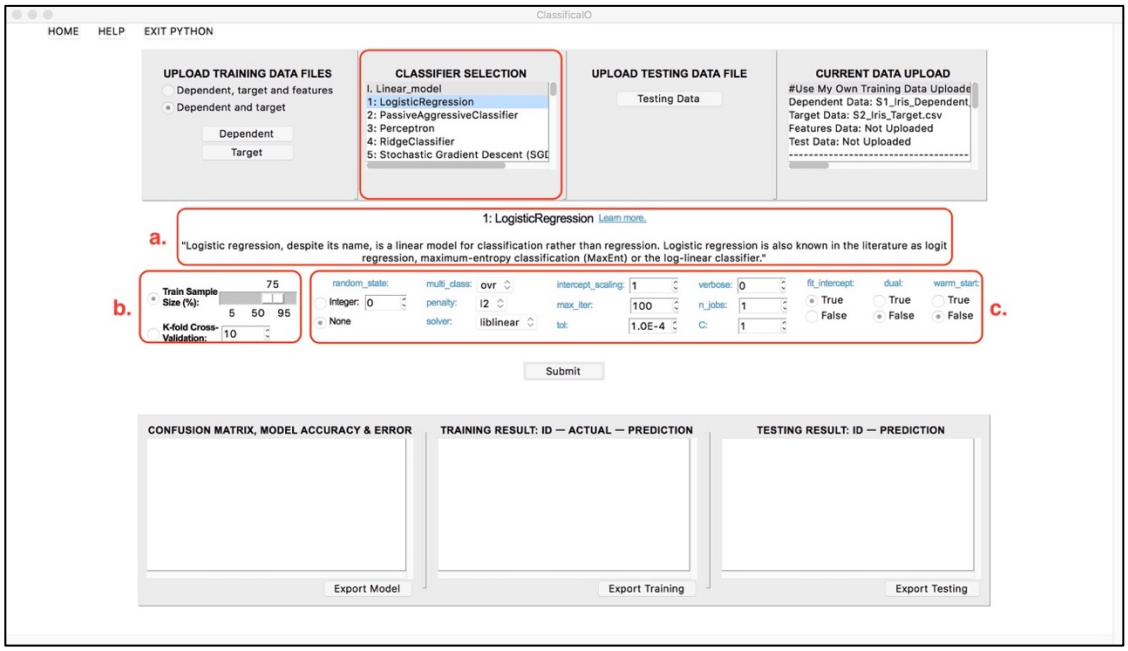
b



	A	B
1	id	target
2	1	0
3	2	0
4	5	0
5	6	0
6	7	0
7	11	0
8	12	0
9	14	0
10	16	0
11	18	0
12	20	0
13	21	0
14	22	0
15	23	0
16	24	0

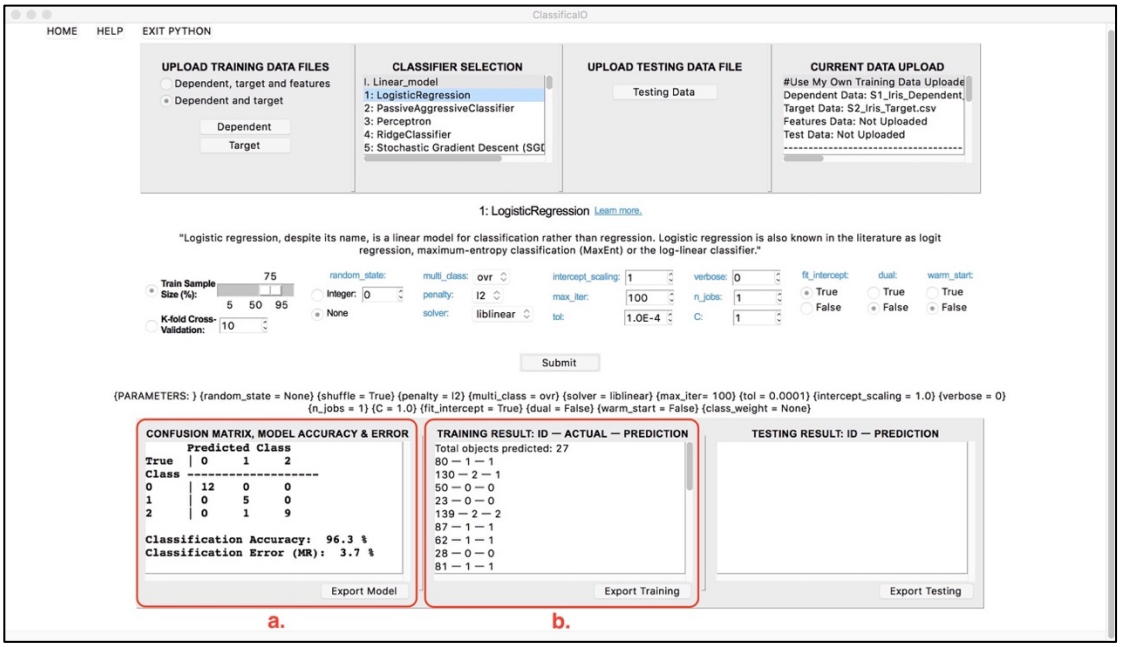
a. Dependent data, example of partial dependent data file format. Testing data (not shown) uses the same format. **b.** Example of partial target data file format where the targets correspond to setosa = 0, versicolor = 1, and virginica = 2. Versicolor and virginica are not visible in this screenshot.

Figure 8. Selected logistic regression classifier.



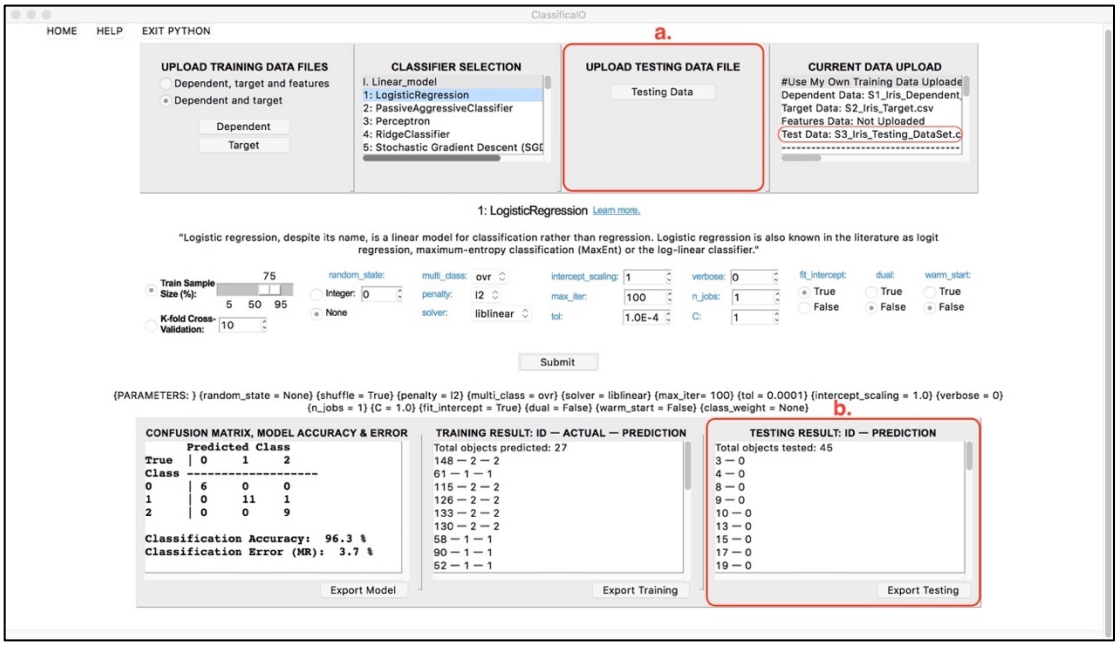
The interface for each selected classifier, has uniform features. **a.** Classifier definition is displayed, together with an underlined clickable link that reads “Learn more” next to the classifier name. **b.** Training methods with ‘Train Sample Size (%)’ method selected. **c.** The classifier parameters set to their default values.

Figure 9. Trained logistic regression classifier.



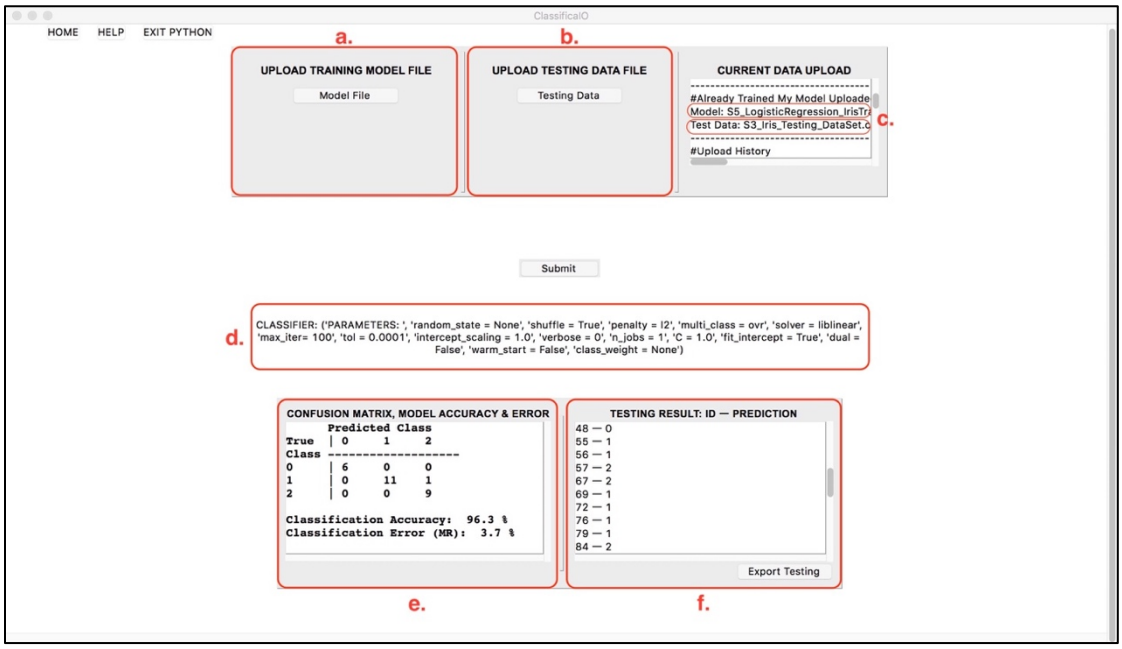
a. Trained model using 78 data points (75% of 105 data points), classifier evaluation (confusion matrix, model accuracy and error). **b.** Model validated using 27 data points (25% of 105 data points).

Figure 10. Tested logistic regression classifier.



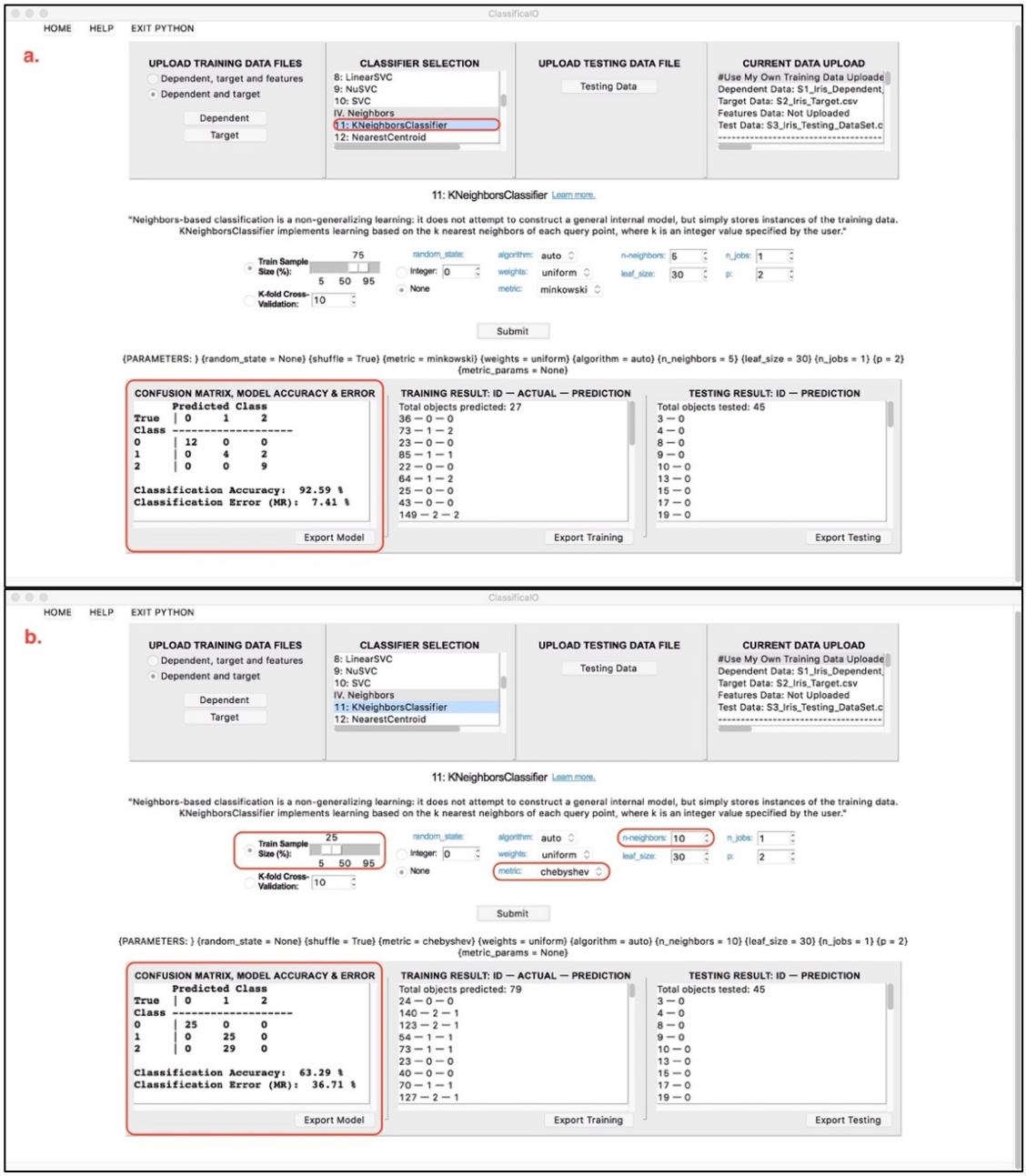
a. Upload testing data panel. **b.** Model tested using 45 data points.

Figaure 11. 'Already Trained My Model' window.



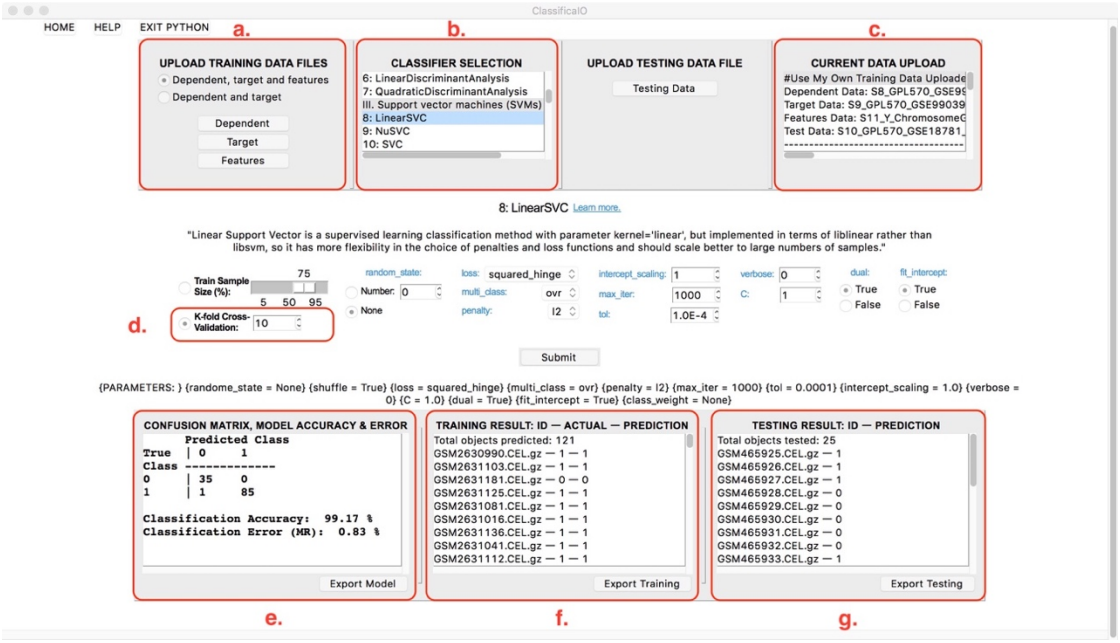
a. Upload ClassificaIO trained model panel. **b.** Upload testing data panel. **c.** Current data upload panel with both model and testing data files names shown (red boxes). **d.** Model preset parameters. **e.** Trained model result and model evaluation (confusion matrix, model accuracy and error). **f.** Model testing result.

Figure 12. Training and testing using gene expression data.



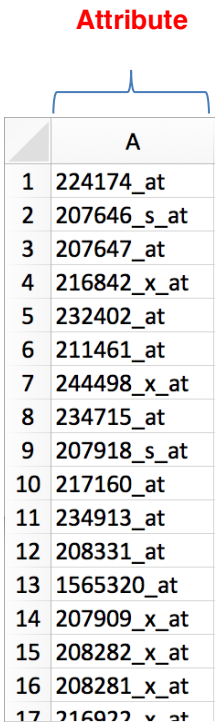
a. selected k-nearest neighbors' classifier with trained and tested the data using the default parameters values, **b.** Same classifier selected with trained and tested data but using different parameters values.

Figure 13. Trained linear support vector machine classifier.



Trained model using GSE99039 121 data points and k-fold cross validation, classifier evaluation (confusion matrix, model accuracy and error). Model validated and tested model using GSE18781 25 data points.

Figure 14. Features data.



Example of partial features data file format where each Affymetrix probe id correspond to a Y chromosome gene.

Table 3. ClassificaIO software information.

Current ClassificaIO Version	1.1.5
Public Links to Executables	PyPI: https://pypi.org/project/ClassificaIO/ GitHub: https://github.com/gmiaslab/ClassificaIO
Distribution License	MIT license (MIT)
Operating Systems	Mac OS X, Linux, and Microsoft Windows
Software Installation Dependencies	Python 3 and Python libraries: Tkinter, Pillow, Pandas, NumPy, scikit-learn and SciPy
Supplementary Data Online Availability	https://github.com/gmiaslab/manuals/tree/master/ClassificaIO/Supplementary%20Files
Contact E-mail	gmiaslab@gmail.com

ClassificaIO is provided as open source software, and distributed on GitHub and PyPI. Up-to-date code, manuals and supplementary example material will be maintained on GitHub.

Table 4. Classification algorithms included in ClassificaIO.

CLASSIFIER	Scikit-learn FUNCTION USED	IMMUTABLE PARAMETERS
1: Logistic regression	LogisticRegression	class_weight = None
2: Passive Aggressive	PassiveAggressiveClassifier	class_weight = None n_iter= None
3: Perceptron	Perceptron	class_weight = None
4: Classifier using Ridge regression	RidgeClassifier	class_weight = None
5: Stochastic Gradient Descent - SGD	SGDClassifier	—
6: Linear Discriminant Analysis	LinearDiscriminantAnalysis	shrinkage= None priors = None
7: Quadratic Discriminant Analysis	QuadraticDiscriminantAnalysis	store_covariances = None priors = None
8: Linear Support Vector	LinearSVC	class_weight = None
9: Nu-Support Vector	NuSVC	class_weight = None
10: C-Support Vector	SVC	class_weight = None
11: k-Nearest Neighbors	KNeighborsClassifier	metric_params = None
12: Nearest centroid	NearestCentroid	—
13: Radius Nearest Neighbors	RadiusNeighborsClassifier	metric_params = None
14: Gaussian Process Classification (GPC)	GaussianProcessClassifier	kernel = None
15: Naive Bayes for Multivariate Bernoulli Models	BernoulliNB	class_prior = None
16: Gaussian Naive Bayes	GaussianNB	class_prior = None
17: Naive Bayes for Multinomial Models	MultinomialNB	class_prior = None
18: Decision Tree	DecisionTreeClassifier	class_weight = None
19: Extremely Randomized Tree	ExtraTreeClassifier	min_impurity_split = None class_weight = None
20: AdaBoost	AdaBoostClassifier	base_estimator = None
21: Bagging	BaggingClassifier	base_estimator = None
22: Extra Trees	ExtraTreesClassifier	class_weight = None
23: Random Forest	RandomForestClassifier	class_weight = None
24: Label Propagation	LabelPropagation	—
25: Neural network Multi-layer Perceptron	MLPClassifier	—

A list of all 25 classification algorithms, their corresponding scikit-learn functions, and immutable (unchangeable) parameters with their default values.

BIBLIOGRAPHY

BIBLIOGRAPHY

1 Mias, G. I. & Snyder, M. Personal genomes, quantitative dynamic omics and personalized medicine. *Quant Biol* **1**, 71-90, doi:10.1007/s40484-013-0005-3 (2013).

2 Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321-332, doi:10.1038/nrg3920 (2015).

3 Buckberry, S., Bent, S. J., Bianco-Miotto, T. & Roberts, C. T. massiR: a method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics* **30**, 2084-2085, doi:10.1093/bioinformatics/btu161 (2014).

4 Ohler, U., Liao, G. C., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol* **3**, RESEARCH0087 (2002).

5 Degroeve, S., De Baets, B., Van de Peer, Y. & Rouze, P. Feature subset selection for splice site prediction. *Bioinformatics* **18 Suppl 2**, S75-83 (2002).

6 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318, doi:10.1038/ng1966 (2007).

7 Way, G. P. *et al.* A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma. *BMC Genomics* **18**, 127, doi:10.1186/s12864-017-3519-7 (2017).

8 Lee, S. I. *et al.* A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun* **9**, 42, doi:10.1038/s41467-017-02465-5 (2018).

9 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).

10 Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* **15**, doi:10.1098/rsif.2017.0387 (2018).

11 Berthold, M. R. *et al.* KNIME: The Konstanz Information Miner. *Stud Class Data Anal*, 319-326, doi:Doi 10.1007/978-3-540-78246-9_38 (2008).

12 Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479-2481, doi:10.1093/bioinformatics/bth261 (2004).

13 Demsar, J. *et al.* Orange: Data Mining Toolbox in Python. *J Mach Learn Res* **14**, 2349-2353 (2013).

14 Scikit Learn Documentation. Scikit learn online documentation. (2018).

15 Help, K. D. a. *How to use kaggle*, <https://www.kaggle.com/docs> (2018).

16 Stack Overflow. The stack overflow python online comunity. (2018).

17 Ousterhout, J. K. *Tcl and the Tk toolkit*. (Addison-Wesley, 1994).

18 McKinney, W. in *Proceedings of the 9th Python in Science Conference.* 51-56.

19 Oliphant, T. E. *A guide to NumPy*. Vol. 1 (Trelgol Publishing USA, 2006).

20 Olivier, B. G., Rohwer, J. M. & Hofmeyr, J. H. S. Modelling cellular processes with Python and Scipy. *Mol Biol Rep* **29**, 249-254, doi:Doi 10.1023/A:1020346417223 (2002).

21 Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann Eugenic* **7**, 179-188, doi:DOI 10.1111/j.1469-1809.1936.tb02137.x (1936).

22 Anderson, E. The Irises of the Gaspe peninsula. *Bulletin of American Iris Society* **59**, 2-5 (1935).

23 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210 (2002).

24 Shamir, R. *et al.* Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **89**, 1676-1683, doi:10.1212/WNL.0000000000004516 (2017).

25 Sharma, S. M. *et al.* Insights in to the pathogenesis of axial spondyloarthropathy based on gene expression profiles. *Arthritis Res Ther* **11**, R168, doi:10.1186/ar2855 (2009).

Chapter 3 –

Computational Meta-analysis of Gene Expression in Acute Myeloid Leukemia

Abstract

In 2018 alone, an estimated 20,000 new acute myeloid leukemia (AML) patients were diagnosed, in the United States, and over 10,000 of them are expected to die from the disease. AML is primarily diagnosed among the elderly (median 68 years old at diagnosis). Prognoses have significantly improved for younger patients, but in patients older than 60 years old as much as 70% of patients will die within a year of diagnosis. In this study, we conducted stratified computational meta-analysis of 2,213 acute myeloid leukemia patients compared to 548 healthy individuals, using curated publicly available data. We carried out analysis of variance of normalized batch corrected data, including considerations for disease, age, tissue and sex. We identified 964 DE unique genes (974 DEPS) differentially expressed genes and 4 associated significant pathways involved in AML. Additionally, we have identified 70 DEPS with 69 unique sex- and 372 age-dependent DE gene signatures relevant to AML. Finally, we used a machine learning model (KNN model) to classify AML patients compared to healthy individuals with > 90% achieved accuracy. Overall our findings provide a new reanalysis of public datasets, that enabled the identification of potential new gene sets relevant to AML that can potentially be used in future experiments and possible stratified disease diagnostics.

A version of this chapter and results has been submitted for publication.

Introduction

Acute myeloid leukemia (AML) is a heterogeneous malignant disease of the hematopoietic system myeloid cell lineage. AML is best characterized by the terminal differentiation in normal blood cells and excessive production and release of cells at various stages of incomplete maturation (leukemia cells). As a result of this faster than normal and uncontrolled growth of leukemia cells, healthy myeloid precursors involved in hematopoiesis are suppressed, and ultimately, can soar to death within months from diagnosis if untreated[1,2]. AML accounts for 70% of myeloid leukemia and nearly 80% of acute leukemia cases, making it the most common form of both myeloid and acute leukemia[2,3]. The number of new AML cases is increasing each year – in 2018 alone, there have been an estimated about 20,000 new diagnosed AML patients, over 10,000 of them will die from the disease[4].

AML can occur in people of all ages but is primarily diagnosed among the elderly (>60 years), with a median age of 68 year at diagnosis[4]. Recent advances in AML biology expanded our understanding of its complex genetic landscape and led to significant improvement in prognoses and therapeutic strategy for younger patients[5,6]. However, in patients older than 60 years old, prognoses remain grim and therapeutic strategy has been nearly the same for more than 30 years[2,5-8]. Approximately 70% of patients 65 years of age or older die within one year from diagnosis[9]. While it is apparent that the nature of AML changes with age, still little is known about the extent of these associations and how they vary with patient's age[6,10,11]. Taking into consideration age considerations in

the identification of changes in AML global gene expression can lead to improved early diagnosis and improvement in treatment approaches for elderly patients.

AML prognosis is highly dependent on cytogenetic analysis since chromosome abnormalities including translocations, deletion, duplication and inversions occur frequently in AML [6]. Cytogenetic analysis, as a prognostic approach, is used to classify AML patients that carry distinctive chromosomal abnormality into either favorable, intermediate, or unfavorable risk group. However, approximately 50% of AML patients lack genetic abnormalities and present normal karyotype[12-15]. In the last decade, many frequently mutated genes in AML were identified including NPM1, CEBPA, RUNX1, FLT3 -- and many studies have reported sets of genes and gene panels that can be used to improve the prognostication of AML[8,12,13]. However, the impact of these findings to help improve AML prognosis in the current clinical practice is still unclear[8].

Gene expression profiling is a powerful prognostic method for the detection of changes in gene expression due to genetic abnormalities, gene fusion and/or mutations in AML patients. In the past, gene expression biomarkers were used to classify myeloid leukemia as compared to lymphoid leukemia including many subtypes within each of the two diseases[16-18]. Multiple gene expression analyses of AML have been carried out, 25 of these have been systematically compared by Miller and Stamatoyannopoulos[19], who analyzed information on 4,918 genes, and identified 25 genes reported across multiple, with potential prognostic features. In this study, we performed comprehensive meta-analysis of 2,213 acute myeloid leukemia patients and 548 healthy subjects using 34

publicly available gene expression microarray datasets (following strict inclusion criteria) to identify disease, sex- and age-related gene expression changes and signaling pathways associated with AML. We identified sex- and age-related gene expression signatures that show similar alteration in gene expression levels and associated signaling pathways in AML and have used our results (gene sets) to predict AML or healthy status. We believe that our results may lead to improved AML early detection and diagnostic testing with target genes, as well as the identification of new targets for treatment with mechanisms of action different from those used in conventional chemotherapy. To our knowledge, provided the body of published AML gene expression studies, our approach of joining multi-study gene expression datasets for meta-analysis to identify disease-, sex- and age-related signatures in AML has not been implemented before.

Results

- **Data curation and gene expression preprocessing**

By navigating the Gene Expression Omnibus (GEO) public repositories according to our systematic workflow and inclusion criteria (Fig. 15a&b), 34 age-annotated gene expression datasets from 32 different studies covering 2,213 AML patients and 548 healthy individuals were curated and selected for gene expression meta-analysis and functional pathway enrichment analysis. Table 5 provides a description on each dataset with a sub-table summary of all curated data used in our current study. After pre-processing each individual data set separately according to Figure 15b, we performed analysis on 44,754 probe sets which were common across all samples (arrays).

- **Classification of missing metadata annotation**

After the data curation step, 805 arrays (802 AML and 3 healthy) of 2,761 curated data were missing sex annotation, and 737 arrays (all AML patients) were missing information regarding sample source (i.e. tissue, either bone marrow [BM] or peripheral blood [PB] annotation). Classification for the missing annotations for these arrays (1,542 in total) was essential in our study, to increase the sample size, and statistical power [20]. To predict the missing sex and sample source meta-data, we trained and validated various machine learning supervised models, including logistic regression (LR) and k-nearest neighbor (KNN) classification models. The models were trained and verified using our annotated preprocessed expression data. Model training, parameters used in training, validation for this analysis are discussed in the method section. Results from model training, including confusion matrix, model accuracy, and error can be viewed in Supplementary Table S1 online and results from classification for missing annotation are presented in Supplement file 1&2.

- **Batch correction**

Our pre-processed data, AML and healthy, was subjected to "dataset-wise correction" (for more details and further explanation, see method section, presented in sub-section **"Dataset-wise correction for batch effects".** We used ComBat[21] to correct for confounding batch effects. Our datasets used in this study did not include within-study healthy controls, which would limit variance analysis, and the ability to separate biological from batch effects. To address this, we implemented an iterative batch effect correction approach, essentially employing a weight-based method for correcting batch

effects. Assuming the batch effects due to each data set is a function of the number of samples in the data set (weight), normalizing sets of unevenly sized datasets may lead to unbalanced batch correction. We used 5 additional datasets as a reference set, which we refer to as "covariate" hereafter. Each of the covariate datasets included within study healthy controls. All 5 datasets together consisted of a total 613 arrays (455 AML and 158 healthy) (Table 5), and pre-processed exactly as our curated data sets. These were used together with each of the remaining datasets to batch correct each dataset with respect the covariate reference using ComBat[22]. After this dataset-wise correction, covariate datasets were removed, and our expression data were clustered using principal component analysis (PCA) to visually examine the effect of covariate datasets on distributing the batch weight during batch correction. Figure 16 shows the results for batch effects correction, including batch corrected data without covariate datasets (Fig. 16 a&b), as well as batch corrected data with covariate datasets (Fig. 16 c&d).

- **Analysis 1: Gene expression meta-analysis and enrichment analysis of AML disease state compared to healthy individuals**

  - **Gene expression meta-analysis of AML disease state**

Following batch correction, we performed an analysis of differential gene expression (DGE) on 34 data sets including 2,213 AML patients and 548 healthy controls. Analysis of Variance (ANOVA)[23-25] was performed according to a linear model (see method section "**Meta-analysis"**). 974 Statistically significant differentially expressed probe sets (DEPS) (with genes corresponding to 964 unique gene symbols) for AML versus healthy

were selected based on a Bonferroni[26] adjusted p-value < 0.01 (accounting for multiple hypothesis testing), in conjunction with a two-tailed 5% quantile selection[27] based on the mean difference distribution between AML-healthy group comparisons (post-hoc analyses using Tukey's Honestly Significant Difference (HSD)). The heatmap (Fig. 17a) shows the gene expression with hierarchical clustering of the 974 DEPS, including 487 up- and 487 down-regulated with respect to AML as compared to healthy. The clustering did not reveal any sub-clustering or structure indicative of a grouping or possible classification in the AML subjects (that would also be suggestive of necessary additional blocking design for a per-class analysis). From this analysis, WT1 (Wilms tumor 1) with mean difference of 0.26 and adjusted p-value < 4.11E-11 was the most DE up-regulated gene while CRISP3 (cysteine-rich secretory protein 3) with mean difference of -0.52 and adjusted p-value < 4.11E-11 was the least DE gene. Figure 17b shows the top 10 up- and down-regulated DEPS with corresponding gene symbols, that resulted from this analysis (also listed in Table 6, including mean difference and Bonferroni p-adjusted values from post-hoc analysis using Tukey's HSD tests). The entire list of all 974 DEPS can be found as Supplementary Table S2 online.

- **Gene enrichment analysis AML disease state DE genes.**

To identify signaling pathways associated DEPS in AML, gene enrichment analysis was performed on all 974 DEPS combined. Signaling pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG)[28-30] and Gene Ontology (GO) terms[31,32] were analyzed for over-representation analysis of biological function in Database for Annotation, Visualization and Integrated Discovery (DAVID)[33,34]. Using Benjamini and Hochberg[35]

adjusted p-value < 0.05, 4 KEGG signaling pathways were identified, including Hematopoietic cell lineage, Cell cycle, p53 signaling pathway, and Transcriptional misregulation in cancer (Fig. 17c). The 4 KEGG signaling pathways and associated DE genes are summarized in Table 7, including unadjusted p-values and Benjamini and Hochberg[35] adjusted p-values. 56 DEPS including 27 up- and 29 down-regulated were associated with these signaling pathways, and the heatmap of their mean differences is shown in Figure 17d. Additionally, 61 DEPS were enriched by 6 other KEGG signaling pathways known to be involved in AML (Fig. 17e). From our gene enrichment analysis for overrepresented biological GO terms, 21 GO terms were statistically significant with 727 DEPS (335 up- and 392 down-regulated). GO terms included protein and microtubule binding for the molecular function (MF) category, inflammatory and immune responses, mitotic nuclear division, and cell proliferation response for the biological process (BP) category, and finally, cytoplasm, extracellular exosome, cytosol, extracellular space, integral component of plasma membrane immune response, and others, for the cellular component (CC) category (Fig. 17f). The entire list of our enrichment analysis results (statistically significant over-representation in KEGG and GO terms) can be found as Supplementary Table S3 online.

- **Analysis 2: gene expression meta-analysis and enrichment analysis of sex- and age-related DE genes in AML**

Further analysis of gene expression and pathways enrichment were conducted in order to characterize sex- and age-specific gene expression changes in AML patients compared to healthy individuals, **Analysis 2a:** "**Sex-relevance differential gene expression meta-**

analysis and associated signaling pathways in AML", and **Analysis 2b:** "**Age-dependent differential gene expression meta-analysis and associated signaling pathways in AML"**. We used the same filtering criteria in both analyses as those used in analysis 1 for significant DE genes and signaling pathways between AML patients and healthy controls. In addition, DE genes were regarded to be covariates statistically significantly (up- or down-regulated) for each factor, sex and age, if displayed Bonferroni adjusted p-value from Tukey's HSD < 2.2×10^{-7} (=0.01/(44,754 probe sets used)).

- **Analysis 2a. Sex-relevance differential gene expression meta-analysis and associated signaling pathways in AML**

Gene expression meta-analysis was also used to identify DEPS that show sex relevance with respect to male AML patients as compared to female AML patients. 266 DEPS were regarded statistically significant (p-value < 2.2×10^{-7}). A list of all 266 DEPS (including up- and down-regulated, gene title and symbol, male-female mean difference, and Bonferroni corrected p-value) can be found as Supplementary Table S4 online. 70 DEPS with 69 unique DE genes were found to overlap between analysis 1 (AML disease state) and analysis 2a (Fig. 18a). Figure 18b shows these 70 DEPS with gene symbol annotations, and their mean difference values in the heatmap, which displays differences in significance for a common DEPS in both analyses 1 and 2. The top 10 up- and down regulated DEPS from this analysis are shown in Figure 18c. Figure 18d shows the gene expression heatmap with a hierarchical clustering of the 70 DEPS (rows) on sex and disease state of all 2,213 AML and 548 healthy subjects (columns) indicated by color bars above the heatmap.

For enrichment analysis, we searched for common DEPS between the 70 DEPS from this meta-analysis and the 974 DEPS from AML disease state meta-analysis, for KEGG pathways and GO terms. 4 sex-relevant DE genes were found in 3 different signaling pathways (Table 8), including, (3 up- and 1 down-regulated). Up-regulated genes and pathway memberships included, FLT3 and CD34 in Hematopoietic cell lineage, FLT3 in Transcriptional misregulation in cancer, and PMAIP1 in p53 signaling pathway, and down-regulated gene MS4A1 in Hematopoietic cell lineage (Fig. 18e).

Figure 18f shows GO analysis results, where 15 overrepresented biological GO terms were overlapped, including terms GO:0005615~extracellular space, GO:0006955~immune response, GO:0005515~protein binding, GO:0005819~spindle, and GO:0030496~midbody. The entire list of our enrichment analysis (statistically significant KEGG and GO terms) can be found as Supplementary Table S3.

- **Analysis 2b. Age-dependent differential gene expression meta-analysis and associated signaling pathways in AML**

Here we refer to the "age-group" to indicate AML patients and healthy individuals in the same age range assigned in our study. The subjects were binned in 8 groups: 0-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80-100 years old. From this meta-analysis, 1395 probsets across all age-groups were identified as statistically significant (Bonferroni adjusted p-value < 2.2×10^{-7}) (Supplementary Table S5). From these 372 DE unique age-dependent genes (375 DEPS) were found to overlap with the 964 DE unique genes (974

DEPS) from our AML disease state meta-analysis (Fig. 19a), with 137 up- and 238 down- regulated. The entire list of 375 DEPS can be found as Supplementary Table S6 online. Figure 19b shows the heatmap of the 375 DEPS (rows) across 18 age-groups (columns) that were deemed statistically significant according to Bonferroni adjusted p-value $< 2.2 \times 10^{-7}$ (these age-groups including all 2,761 arrays (2,213 AML patients and 548 healthy individuals)). The top 10 up- and down- regulated DE genes from this analysis are shown in Figure 19c. Figure 19d shows 75 DE genes identified to have appeared specifically in one age-group.

To investigate further, pairwise correlations between age-groups were computed (Fig. 19e). The "0 to 19" age-group was used as a common comparison reference with respect to other groups (Fig. 19f). Using this "0 to 19" group as a baseline, Figure 19g shows the mean difference of 25 genes that are DE with respect to the "0 to 19" baseline across all other groups and the mean difference values between AML and healthy are shown in the right-most column of Figure 19a-g for reference. Utilizing results for KEGG analysis for signaling pathways from analysis 1, Figure 19 shows 17 DE genes identified in all 4 KEGG pathways according to age groups (also listed in Table 9).

- **Age-dependent genes analysis for drug to gene interaction**

We carried out two analysis in The Drug Gene Interaction Database (DGIdb)[36] to identify druggable genes or gene products from our results with known drug to gene interaction (where available) for a potential therapeutic in AML. According to DGIdb, druggable genes are defined as "genes or gene products that are known or predicted to interact with

drugs, ideally with a therapeutic benefit to the patient". The two analyses were performed in DGIdb using two different gene sets, (i) 25 DE genes common across the baseline (0 to 19) age-group (Fig. 19g) and, (ii) 75 genes identified to be specific to one age-group (Fig. 19d). Table 10 lists all the genes and their corresponding associated druggable gene categories.

Discussion

According to the 2016 World Health Organization (WHO) newly revised myeloid neoplasms and acute leukemia classification system[37], AML prognosis criteria for classification is highly dependent on the presence of chromosomal abnormalities, including chromosomal deletions, duplications, translocations, inversions, and gene fusion. Mostly, AML is diagnosed through microscopic, cytogenetics, and molecular genetic analyses of patients' blood and/or bone marrow samples. Microscopic examination is used to detect distinctive features (e.g. Auer rods) in cell morphology, cytogenetic analysis to identify chromosomal structural aberrations (e.g., t(8;21), inv(16), t(16;16), or t(9;11)), and molecular genetic analysis to identify gene fusion (e.g., RUNX1-RUNX1T1 and CBFB-MYH11), and mutations in genes frequently mutated in AML (e.g., NPM1, CEBPA, RUNX1, FLT3)[8,12,13]. These cytogenetic and molecular genetic analyses are used to identify prognosis markers that can be used to classify AML patients into three risk categories: favorable, intermediate, and unfavorable. The largest group of AML patients (almost 50%) however, present normal karyotype and lack genetic abnormalities[12-15]. These patients are classified as intermediate risk, and often have heterogeneous clinical outcome with standard therapy with risk of AML relaps[38].

Additionally, AML prognosis worsens as age increases, and older patients respond less to current treatments with poorer clinical outcomes than their younger counterparts[5,39]. Further complicating, AML has multiple driver mutations and competing clones that evolve over time, making it a very dynamic disease[40,41]. Identifying differentially expressed genes and associated signaling pathways based on our analysis, that incorporates disease state, sex- and age-dependent meta-analysis, can provide global gene expression signatures, which collectively can potentially serve as sex- and age-dependent biomarkers for AML prognosis compared to healthy.

In the present study, we aimed to establish, disease sex-linked and age-dependent biomarkers from genes with similar alteration in gene expression level and associated signaling pathways in AML. Utilizing microarray gene expression data and combined with various machine learning models, respectively, our biomarkers were indicative of prognostic signature for AML prediction compared to healthy with > 90% achieved accuracy. We took advantage of 34 publicly available microarray gene expression data sets covering 2213 AML patients and 548 healthy individuals to identify changes in AML gene expression associated with disease state (AML compared to healthy), sex-linked (male compared to female), and age-dependent (across age-groups compared to baseline). We performed 3 differential gene expression and gene enrichment analyses:

Analysis 1: Gene expression meta-analysis and associated signaling pathways of AML disease state compared to healthy individuals, was carried out to identify DE genes in AML disease state, followed by gene enrichment analysis on the identified DE

genes to find singling pathway associated with AML. The results from this analysis were then used as baseline indicator for AML disease state.

Analysis 2a: Sex-dependent gene expression meta-analysis and associated signaling pathways in AML compared to healthy individuals, was performed to explore the relevance of patients' sex on gene expression and to identify sex-linked genes and associated signaling pathways in AML.

Analysis 2b: Age-dependent gene expression meta-analysis and associated signaling pathways in AML compared to healthy individuals, was carried out to identify common set of age-dependent genes and associated signaling pathways and to explore age-dependent trends in gene expression in AML.

- **Analysis 1 discussion: Gene expression meta-analysis of AML disease state**

From our meta-analysis for AML disease state 964 DE unique genes (974 DEPS) (487 overexpressed and 487 underexpressed) were identified as significantly differentially expressed between AML patients and healthy individuals (Bonferroni adjusted p-value < 0.01). Among these 6 genes are known to be involved in AML functional pathways, including 4 up-regulated, JUP (junction plakoglobin), CCNA1 (cyclin A1), FLT3 (fms-related tyrosine kinase 3), PIK3R1 (phosphoinositide-3-kinase, regulatory subunit 1 (alpha)), and 2 down-regulated, CD14 (CD14 molecule), CEBPE (CCAAT/enhancer binding protein (C/EBP), epsilon). The top 10 up- and down-regulated genes from this analysis are listed in Table 6 with their respected Tukey's HSD mean difference and

Bonferroni p-adjusted values. As shown in Figure 17b of the top 10 up- and down-regulated DEPS -- WT1 (Wilms tumor 1) was found to be the most expressed and CRISP3 (cysteine-rich secretory protein 3) was the most under-expressed gene. From our gene enrichment analysis for overrepresented biological GO terms, WT1 is identified with protein binding and cytoplasm in the molecular function (MF) and cellular component (CC) categories respectively. WT1 is a transcriptional regulatory protein essential to cellular development and cell survival, and it has been known to be highly expressed with an oncogenic role in AML[42,43], in agreement with our findings. However, CRISP3's direct role in AML is still under investigation. CRISP3 is a member of the cysteine-rich secretory protein CRISP family with major role in female and male reproductive tract, and is mainly expressed in salivary gland and bone marrow[44]. Recently in 2017, 80 genes were reported as "extracellular matrix specific genes" in leukemia, and CRISP3 was among the downregulated DE genes reported.[45] In GO terms from our gene enrichment analysis, CRISP3 is associated with the extracellular space, specific granule, and extracellular exosome GO terms, all in the cellular component (CC) category. These findings suggest that CRISP3 could be a potential candidate as a prognostic biomarker in AML. CRISP3 associations with these cellular components in AML have not been previously reported, to the best of our knowledge and merit further investigation.

The enrichment analysis for overrepresented biological GO terms of the 974 DEPS (up- and down-regulated combined) is shown in Figure 17f. 727 DEPS (335 up- and 392 down-regulated) were enriched for 21 GO terms. 592 of which (257 up- and 335 down-

regulated) were enriched in the cellular component (CC) category and were mainly associated with cytoplasm, extracellular exosome, cytosol, and extracellular space. Possible explanations to the relatively high number of DEPS associated with these GO terms might reflect the bone marrow or immunosuppressive microenvironment which is inevitable to AML development and progression[46,47]. On the biological process (BP) category, GO term were associated with inflammatory and immune responses, and cell proliferation. This is reflective of AML characteristics. AML is characterized by terminal differentiation of normal blood cells and excessive proliferation and release of abnormally differentiated myeloid cells. This faster than normal cell proliferation and uncontrolled growth leads to accumulation of genetic abnormalities that very likely can affect many signaling pathways essential to the immune system.

Figure 17c shows the four statistically significant KEGG pathways identified in the pathway enrichment analysis with the number of DEPS enriched by each pathway, which encompassed 56 DE unique genes (Table 7). Specifically, Figure 17c indicates that Transcriptional misregulation in cancer was the most up-regulated pathway in AML (13 up-regulated DE genes), while Hematopoietic cell lineage, and Cell cycle pathways were mostly down-regulated, and the p53 signaling pathway was balanced in terms of up/down-regulated DE genes. The enriched pathways Figure 17e shows the mean difference values of the 56 DE pathway-associated genes, including 27 genes up- and 29 down-regulated. 61 DEPS from our AML disease state meta-analysis were also associated by 6 other KEGG signaling pathways that are known to be involved in AML (Fig. 17e). All these 10 KEGG pathways are known to be involved in tumorigenesis.

Additionally, the majority of the associated DE genes from AML meta-analysis with the identified signaling pathways are known to be abnormally expressed in AML. These findings are consistent with findings from other studies and our current understanding of AML pathogenesis.

- **Analysis 2a discussion**

To identify DE genes associated with sex in AML, we used post-hoc Tukey's HSD tests for comparison between male and female subjects. A total of 266 genes were found statistically significant in this analysis. 70 of there were also found to overlap with the DE genes from analysis 1, AML disease state meta-analysis (Fig 18a&b). Figure 18c shows the top 10 up- and down-regulated DE genes with respect sex – DDX3Y (DEAD-Box Helicase 3 Y-Linked), EIF1AY (Eukaryotic Translation Initiation Factor 1A Y-Linked), KDM5D (Lysine Demethylase 5D), RPS4Y1 (Ribosomal Protein S4 Y-Linked 1) were the most expressed genes and XIST (X Inactive Specific Transcript), TSIX (TSIX Transcript, XIST Antisense RNA), and PRKX (Protein Kinase X-Linked) were the most down-regulated genes. These genes are known to be sex-specific and show such differences and sex separation within the AML and the healthy groups respectively (Fig. 18d). The role of these genes as positive controls in studies with AML needs to be investigated further. We also reported sex and AML known genes that were statistically significant in our analysis, including FLT3 and MAL.

- **Analysis 2b discussion**

The age-dependent meta-analysis in AML using ANOVA, identified 1,381 genes as statistically significant based on Bonferroni adjusted p-value <0.01. We then evaluated the overlap of DE genes from this analysis to our findings of 964 DE unique genes (974 DEPS) in AML disease state (analysis 1) to identify age-related DE genes in AML (Fig. 19a). We identified an overlap of 372 DE unique age-dependent genes (375 DEPS), including 137 up- and 238-down regulated, of age-related genes and AML-associated genes (Bonferroni adjusted p.value <0.01). As shown in Figure 19c, the top 10 most and least expressed age-associate genes in AML according to the mean difference values conducted using Tukey's HSD in seven age-groups, including their corresponding values from AML disease state in column "AML - healthy" for comparisons. Interestingly, CRISP3 (cysteine-rich secretory protein 3) was among the down regulated genes specifically associated with younger age groups, 20 to 49 years of age as compared to 0 to 19 years old. These finding providing our previous finding, are suggestive of CRISP3's role in AML as well as association with certain age-groups. The Figure 19b also shows a number of up-regulated genes known to be involved in AML, including HOXA3, HOXA5 and HOXA10-HOXA9, which belong to the homeobox genes (HOX) family of transcription factors, essential to embryonic development and hematopoiesis, and associated with chromosomal abnormalities translocation and over-expression in AML[48,49]. Interestingly, ORM1 (Orosomucoid 1), which was deemed significant for down-regulation for both of our previous analyses. In fact, in analysis 1, ORM1 was among the top-10 most underexpressed genes, and was also among the 70 DEPS from analysis 2a. These results suggest that ORM1 role in AML is independent of sex or age.

ORM1's direct role in AML also merit's further investigation, given ORM1 involvement in immunosuppression and inflammation[50].

From the 25 DE genes found across the "0 to 19" age-group as a baseline (Fig. 19g), 15 genes were identified as "druggable genome" from our DGIdb[36] analysis (i), including TFF3, ORM1, CA4, CYP4F2, CYP4F3, CEACAM1, FLT3, CHIT1, OLR1, KCNJ15, CAMP, CRISP2, CAPN3, SLC37A3, FCRL1 (Table 10). From these 15 genes, CA4 (carbonic anhydrase 4) showed an interaction record with drug Topiramate, which is an anticancer drug known to act as an inhibiter to CD4[51]. Additionally, we have identified 75 statistically significantly DE genes that show association with only one age-group, exclusively from all other age-groups, suggestive of potential age-specific DGE. Finally, our DGIdb[36] analysis (ii) results for the 75 DE genes, 24 genes were categorized for "druggable genome" including CDH1, GPX3, CD14, DYRK2, SLPI, CCNA2, TGFBR3, UGCG, FCN1, GZMA, TCN1, BPI, S100A12, CDK6, IL12A, P2RY13, ADGRG3, DNMT3B, GUCY1A3, FGFBP2, PTPRJ, LRRK2, BCL2L15, STYX (Table 10).

In summary, our study successfully integrated multiple datasets to perform a study of gene expression in AML, across multiple factors that included disease, sex and age considerations, and identified interesting genes, both known and not previously reported as differentially expressed in each factor. We identified 964 DE unique genes (974 DEPS) and 4 associated significant pathways involved in AML, and 69 DE unique sex-relevant genes (70 DEPS) and 372 DE unique age-related genes (375 DEPS). Using the 964 DE genes, a KNN model allowed for classification of AML patients with >90%

accuracy. We hope that these findings may provide additional relevant targets for further experimental mechanistic studies, and to help identify new markers and therapeutic targets for AML.

- **Future research possibilities and study limitations**

We note that our study identified multiple potentially significant DE genes, associated with age and sex related differences associated with AML as compared to healthy, as discussed above for analyses 1 and 2. While our results and analyses have identified important expression relevant to AML, and many potential new gene targets, we need to acknowledge the limitations of our data: primarily the analysis of AML and healthy subjects involved bone-marrow and blood samples respectively in each group. We tried to account for this disparity in the tissues, by utilizing tissue (sample source) directly as a factor in our linear model, and including its binary interactions with all other factors. Other limitations included an unbalanced AML/healthy ratio, as well as the lack of in-study healthy controls. To address these, we attempted to account for batch effects using a dataset-wise iterative batch correction transformation, as discussed in the method section, presented in sub-section titled, **"Dataset-wise correction for batch effects"**. Finally, a general limitation of utilizing publicly available data is the lack of uniform annotation: the majority of sample data provided have no information on the chromosomal abnormalities, AML classification, and retrospective outcome information. While we accounted for lack of annotations for sex and tissue information using machine learning, which greatly increased our sample size, we would recommend stricter, more

extensive reporting requirements for metadata of publicly available data, deposited in public databases.

Our findings may generate further data-driven investigations including, i) associations between age-groups and changes in gene expression across different AML subgroups to help improve AML risk stratification, ii) age-dependent pseudo time-series models to identify changes in gene expression with more specific AML patients age and sex however such an analysis would require many more well annotated samples that are currently unavailable, particularly given the heterogeneity of the disease, and we hope new studies will address this in the future. Additionally, the use of microarray data is limiting, in that the transcriptome is not fully probed. The availability of more RNA-sequencing data can address this in future expression analyses, additionally involving considerations of allele-specific expression or alternative splicing. Finally, we hope our study will be a resource for the AML research community, as a starting for new hypothesis-driven investigations, that can further probe the mechanistic details of the genes identified as involved in AML, including their possible use as prognostic markers.

Methods

The generalized workflow consisted of five main steps: i) Curation of microarray gene expression data, ii) Preprocessing of raw data files followed by batch effect correction, iii) Predictions of missing annotation data using supervised machine learning, iv) Differential gene expression analysis, and v) Gene enrichment for pathway analysis that includes gene annotation, and finally gene expression-based prediction of AML (Fig. 15a).

- **Gene expression data curation and screening criteria**

Datasets used in this study were selected from the GEO database, maintained by the National Center for Biotechnology Information (NCBI)[52] (https://www.ncbi.nlm.nih.gov/geo/). GEO is a public database repository at the NCBI that function as a hub for high-throughput gene expression datasets storage and retrieval to promote data sharing between researchers. To facilitate speed of search and keep up-to-date with possible new and relevant datasets, as soon as they were released, a Python script was used that utilized functions from the Entrez Utilities from Biopython[53]. The script navigated GEO public database, and downloaded publicly available microarray gene expression datasets. We additionally utilized Python packages, including Pandas, NumPy, and Matplotlib for data structure, numerical computing for data processing, and data visualization respectively. We used strict inclusion criteria to maintain consistency in each dataset selection, screen for availability of both raw and meta-data annotation files provided, human samples used from untreated subjects, and that the sample source

was from either bone marrow (BM) and/or peripheral blood (PB). Inclusion criteria and the data curation workflow are illustrated in Figure 15a-b.

- **Gene expression data sets used in our analysis**

For our analysis we included 34 age-dependent datasets from 32 different studies, 16 included AML and 18 healthy subjects respectively. From the 34 datasets, 32 were produced from Affymetrix GeneChip Human Genome U133 Plus 2.0 (GPL570) and 2 conducted on Affymetrix GeneChip Human Genome U133 Array Set (GPL96 & GPL97) arrays. Table 5 provides detailed information about each data set, including the number of samples used from each dataset, sample tissue source, as well as the total number of AML patients and healthy subjects. Two studies, GSE12417[54] and GSE37642[55-58], were originally conducted on two different Affymetrix array types (GPL570 and GPL96 & GPL97), so each was separated into two subgroups and each subgroup was considered as individual dataset in our meta-analysis, data set GSE12417: (i) subgroup 1 included 73 BM and 5 PB samples, and (ii) subgroup 2 included 160 BM and 2PB. For dataset GSE37642 (i) subgroup 1 included 140 BM and (ii) subgroup 2 422 BM samples (Table 5).

- **Datasets annotation and preprocessing**

Figure 15b outlines the workflow of our preliminary data analysis including preprocessing. For each dataset used in our analysis, raw microarray CEL files were downloaded from GEO, metadata was reviewed, and the data was manually curated to guarantee that and each array, which corresponded to either an AML patient or healthy

individual, was verified and correctly annotated for sample source (BM or PB), platform technology used, age, sex, and disease state (AML or healthy). Raw CEL files from individual datasets were individually pre-processed using the RMA (Robust Multi-Array Average) algorithm[59-61]. Datasets with mixed sample source, i.e both BM and PB, were pre-processed together irrespective of sample source. Preprocessing consisted of correction for background noise using RMA background correction on perfect match (PM) raw intensities, quantile normalization to obtain the same empirical distribution of intensities for each array, median polish summarization of probes into probe sets to estimate gene-level expression value, and logarithm base-2 transformations of gene expression values to facilitate data interpretation (normal distributions) and comparisons between arrays. Additionally, our expression data were first reduced to 44,754 probe sets that are common to and appeared in all data. Data sets were z-score standardized across all probe sets and arrays. Finally, each pre-processed dataset was visualized with box-whisker plots to ensure similar gene expression data distribution across all datasets.

- **Prediction of missing sex- and sample source annotations from curated datasets**

805 arrays (802 from AML patients and 3 were healthy subjects) of curated data were not annotated for sex, while 737 arrays (all AML patients) were missing sample source information. Without these metadata, we would have to discard the data, which in turn would limit the statistical power for the study, and our ability to correct for biases stemming from individual datasets[20]. To address this, we used supervised machine learning classifiers to predict metadata. For all prediction, we used ClassificaIO[62], a

machine learning for classification user interface, which we recently developed, to carry out the machine learning classification analyses utilizing the sklearn package in Python[63]

To predict sex and sample source, pre-processed data sets, 1956 arrays for 545 healthy and 1411 AML, that include 44,754 probe sets and their annotated sex and sample source information were used to train logistic regression (LR) and k-nearest neighbor (KNN) classification models.

The supervised machine learning LR classifier we used with the following parameters: *random_state = None, shuffle = True, penalty = l2, multi_class = ovr, solver = liblinear, max_iter= 100, tol = 0.0001, intercept_scaling = 1.0, verbose = 0, n_jobs = 1, C = 1.0, fit_intercept = True, dual = False, warm_start = False, class_weight = None*

The trained models for classification of missing sex and sample source annotation from curated data achieved > 95% classification accuracy with ~ 3-5% classification errors. Confusion matrix details, model accuracy and error for training and testing are presented in Supplementary Table S1 online, and results in Supplement file 1&2. To account for training overfitting, we used 10-fold cross-validation on all 1,956 gene expression data arrays for training and validation.

- **Dataset-wise correction approach for batch effects correction**

Batch correction was done using a dataset-wise correction. Here we refer to the term "dataset-wise correction," to indicate performing batch correction iteratively on one

dataset at a time, against a reference set of datasets chosen to account for variability. We used this approach to account for the lack within-study healthy controls in the curated gene expression datasets. To address this issue, we used 5 additional datasets the included within-study controls, GEO accessions: GSE107968, GSE68172[64], GSE17054[65], GSE33223[66], and GSE15061[67] (Table 5). We refer to the latter datasets hereafter as "covariate datasets", as they were as the reference datasets in the batch correction. Our approach aimed to balance/distribute the weight of batch effects exerted by each dataset, as this is dependent on the number of observations within a given dataset. Combined, the covariate datasets included 613 total arrays, totaling 455 AML and 158 healthy controls. We used ComBat[22] to correct for study batch effects, as its empirical Bayes-based algorithm uses both scale and mean center based methods, providing an appropriate algorithm[22]. Covariate datasets were treated as the covariate for batch during batch correction, to improve performance in correcting for batch effects rather than biological variation. After batch correction, we used principal component analysis (PCA), visualizing components in both 2 and 3 dimensions, to compare the clustering results for corrections. Covariate data sets were removed after the batch correction step and were not part of our downstream meta-analysis. (Fig. 16a-d).

- **Gene expression meta-analysis**

After batch correction step, we performed gene expression meta-analysis for differential expression on the merged datasets (34 data sets, 16 AML and 18 healthy), where the expression values for all 44,754 common probe sets were aggregated. The effects of patients' age, sex, and sample source, including their pairwise interactions were

investigated using an analysis of variance (ANOVA)[13,68] . The linear model of probe set *i* is then written as:

For each gene i, where i=[1,…44,754], the gene expression

Probeset Y_i was modeled computationally as a linear model:

$$Y_i \sim (a + s + d + t) + (a:s + a:d + a:t) + (s:d + s:t) + (d:t) + \varepsilon$$

Where *d* is the disease state (AML or healthy), *a* is age (between 0 to 100 years), *s* is sex (female or male), *t* is sample source (BM or PB), and ε is a random error term. We note that the model includes sample source and its interactions to address comparisons involving different tissues in AML and healthy subjects (BM or PB respectively).

From the ANOVA analysis, genes were deemed to be disease state statistically significant (differentially expressed) if they displayed ANOVA Bonferroni-adjusted p-value < 0.01. Post-hoc analysis for significant genes was conducted for comparisons (between groups) using Tukey's Honestly Significant Difference (HSD) tests. Additionally, we performed a quantile-based effect filter, were genes were deemed to show biological effects in our analysis if they displayed mean difference values in the <5% and/or > 95% quantiles of the mean difference distributions of the binary group comparisons. Based on the post-hoc analysis, genes were deemed to be statistically significantly (up- or down-regulated) if they displayed Tukey HSD using a Bonferroni adjusted cutoff for p-value < 0.01/44,754.

- **Functional and pathway enrichment analysis**

We carried our enrichment analysis for differentially expressed genes using the Database DAVID[33,34], the KEGG database[28-30] for signaling pathways, GO terms functional annotation for over representation of biological function [31,32] were utilized and signaling pathways were deemed significant based on Benjamini- Hochberg adjusted p-value < 0.05.

- **Using k-nearest neighbor to predict AML**

Before gene expression data passed to the k-nearest neighbor (KNN) algorithm to train, gene expression signatures resulted from our meta-analysis were used to extract expression values. KNN in ClassificaIO[62] was used to carry out this analysis. All 34 data sets (16 AML and 18 healthy) were used for training, and testing was done on all 5 covariate data sets, include AML and healthy subjects. Dependent, target , and testing data files were prepared in accordance with ClassificaIO[62] user guide. The KNN model used the following parameters:

random_state = None, shuffle = True, metric = minkowski, weights = uniform, algorithm = auto, n_neighbors = 5, leaf_size = 30, n_jobs = 1, p = 2, metric_params = None

- **Online data availability**

Supplementary data, tables, figures and files are available online at https://www.zenodo.org/record/1492796#.XA7iUC3Mw_U.

APPENDIX

APPENDIX

Figure 15. General approach, data curation, and analysis workflow summary.

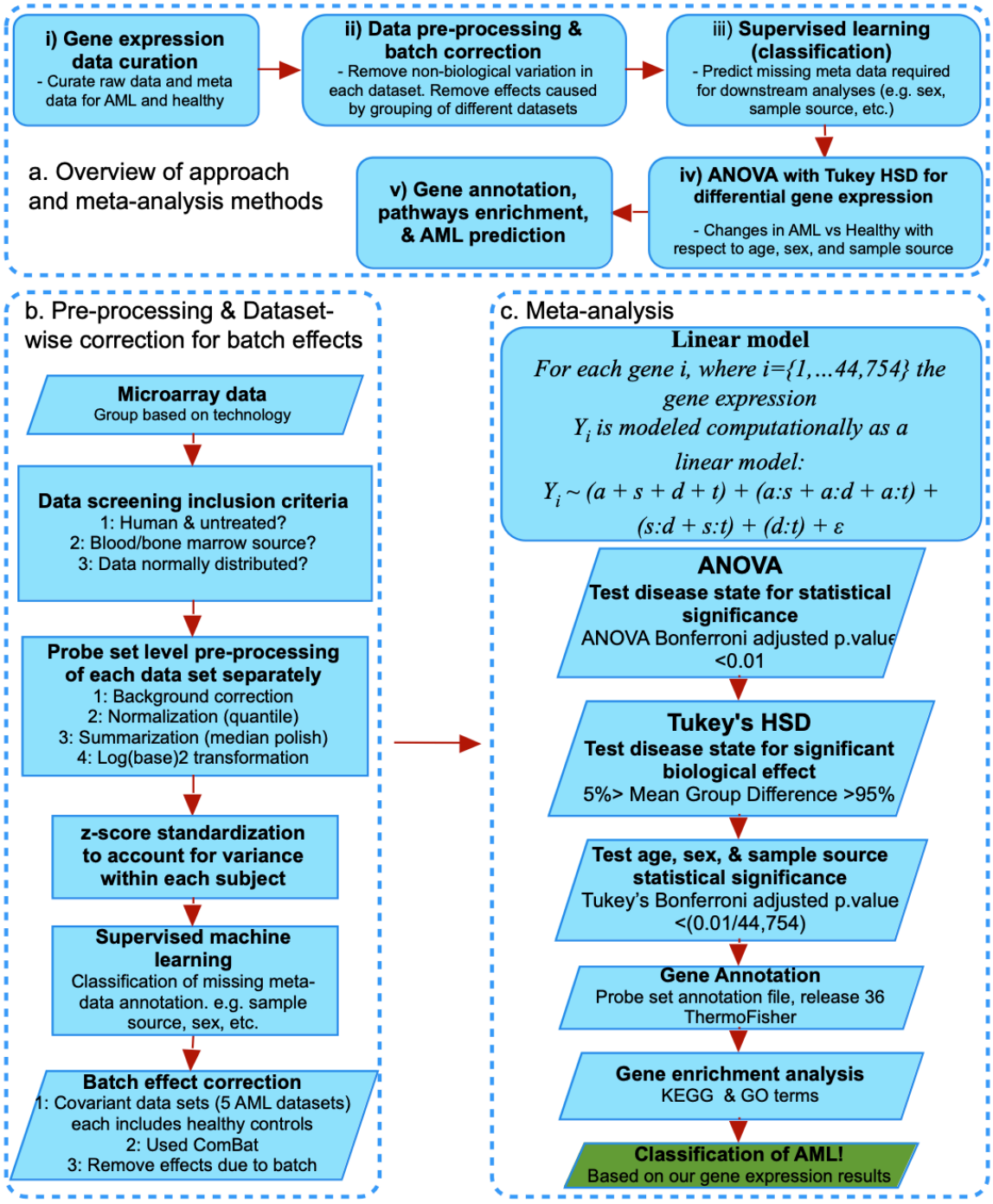


Figure 15. (cont'd)

a. The five main steps that summarize our method of approach for our study. **b.** The curation and screening criteria for raw gene expression and annotation data files curation, data pre-processing, supervised machine learning for missing metadata prediction, and batch effects correction. **c.** Meta-analysis, using linear model in Analysis of Variance (ANOVA) coupled with Post-hoc comparison tested by Tukey's Honestly Significant Difference (HSD), and KEGG enrichment and GO term ontology for signaling pathway and biological function annotations. Finally, classification of AML based on our results.

Figure 16. Principal component analysis of all 2,761 subjects before and after batch correction.

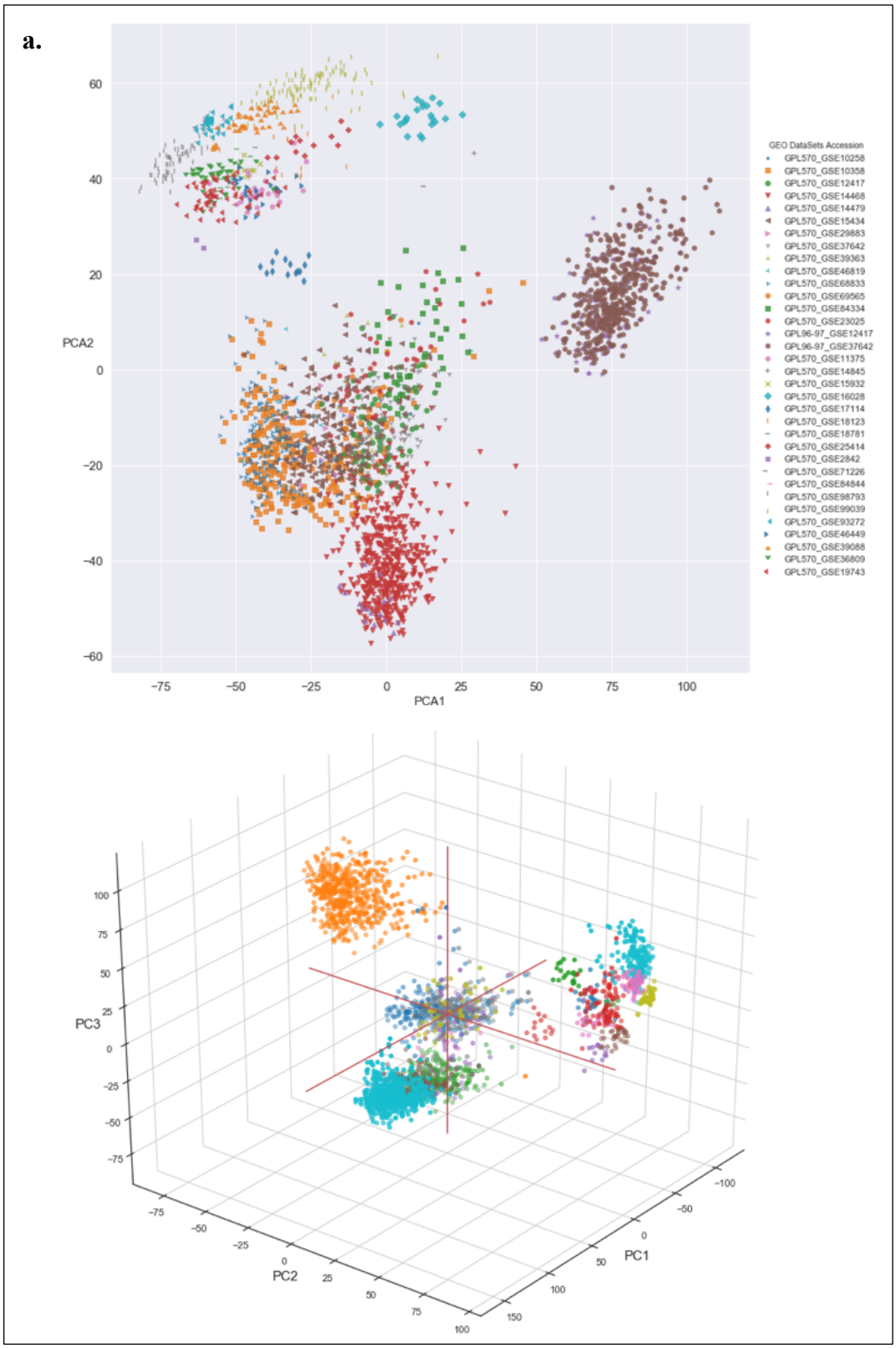


Figure 16 (cont'd)

Figure 16 (cont'd)

c.

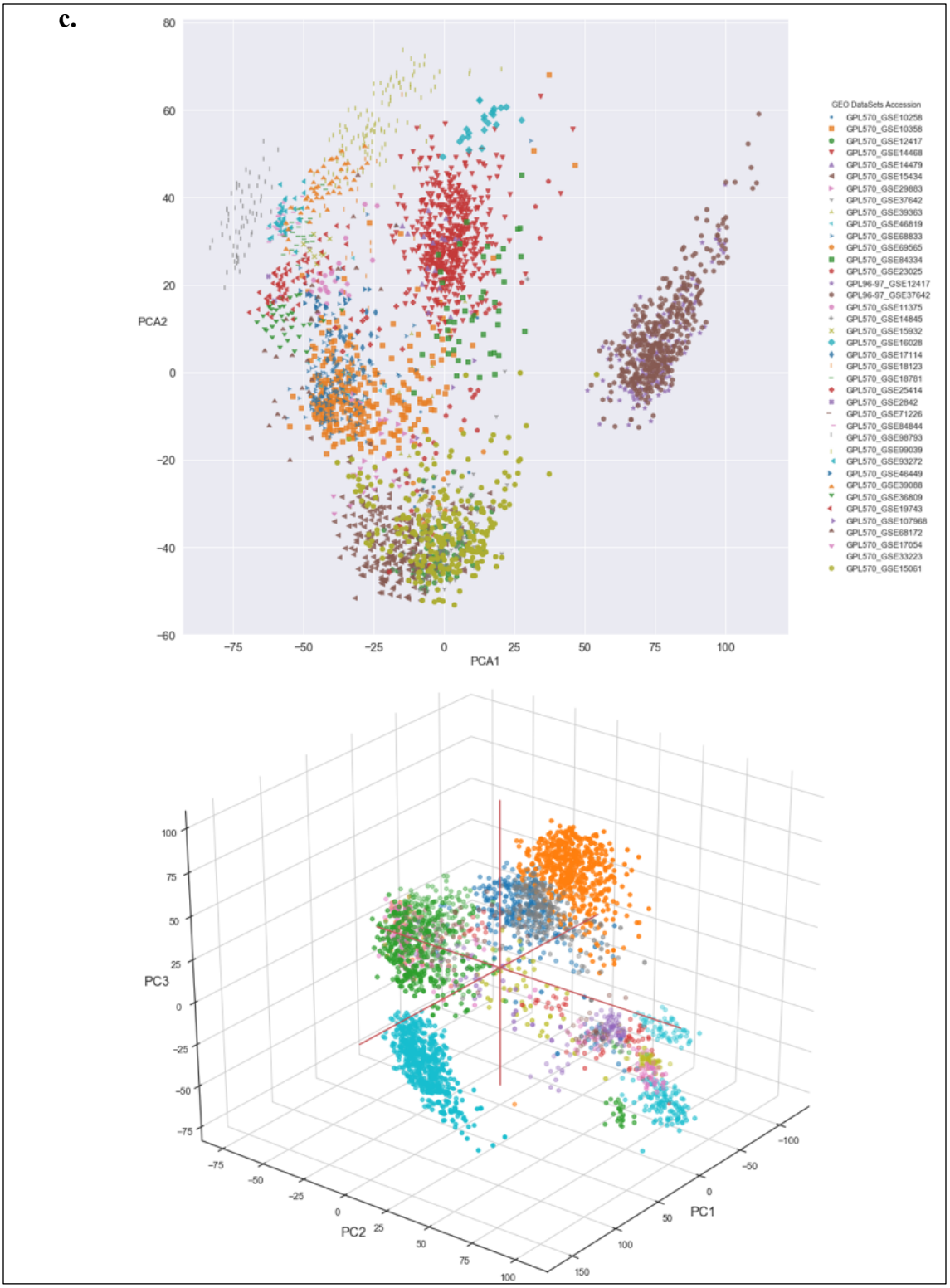


Figure 16 (cont'd)

d.

Figure 16 (cont'd)

For all panels, the first two principal components are at the top and the first three principal components are at the bottom. The data shown in all panels represent gene expression data from 2,761 subjects (2,213 AML patients and 548 healthy individuals) with 44,754 probe sets that has been pre-processed, logarithm base-2 transformed, z-score standardized across all data sets. Panels **a** and **b** show the principal component analysis (PCA) of batch corrected data not including "covariate" datasets, while panels **c** and **d** show the same batch corrected data but including the 5 "covariate" datasets. **a.** Visualizations of the first two and three principal components of gene expression data before batch correction. **b**. The data was corrected without covariate datasets resulting to loss of biological effect information due to lack of within-study controls. **c.** Shows the first two and three principal components of gene expression data including 5 "covariate" data sets (see legend: last 5 labels) that include within-study controls (455 AML and 158 healthy), and **d.** the first two and three principal components of the same data post "dataset-wise correction" for batch effect using ComBat as descripted in the methods section.

Figure 17. Functional classification of DEPS from AML disease state meta-analysis and associated KEGG and GO enrichment analysis.

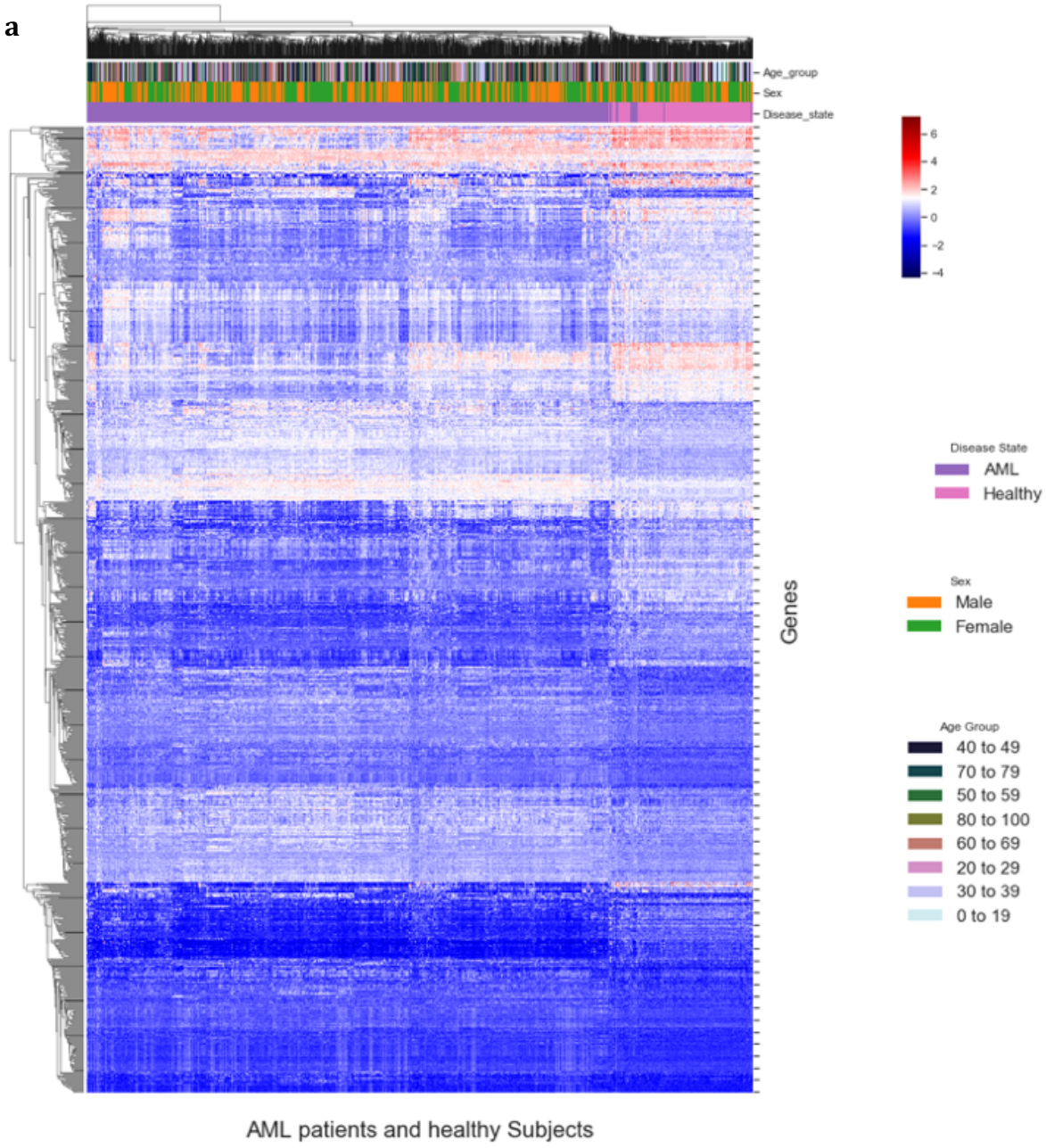


Figure 17 (cont'd)

b

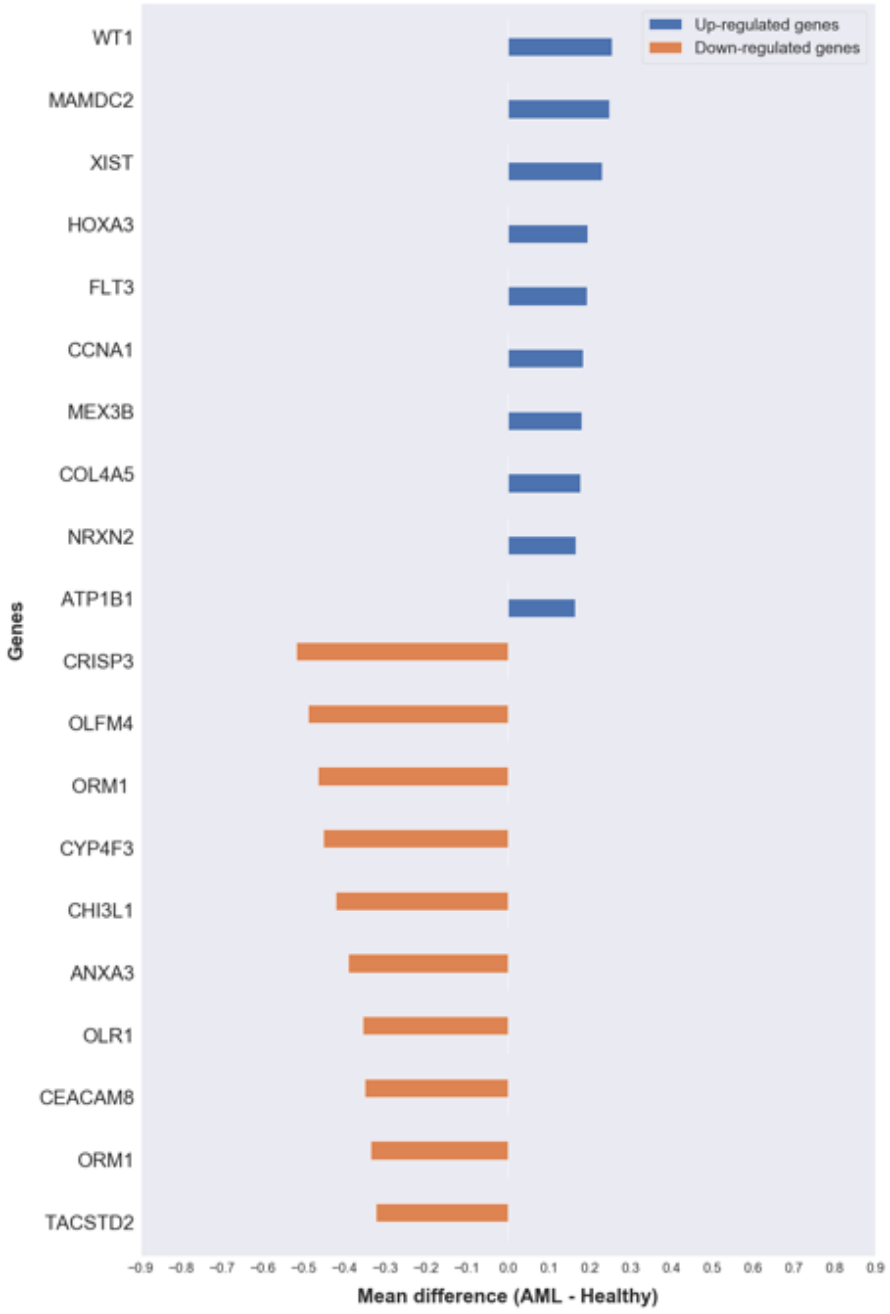


Figure 17 (cont'd)

c

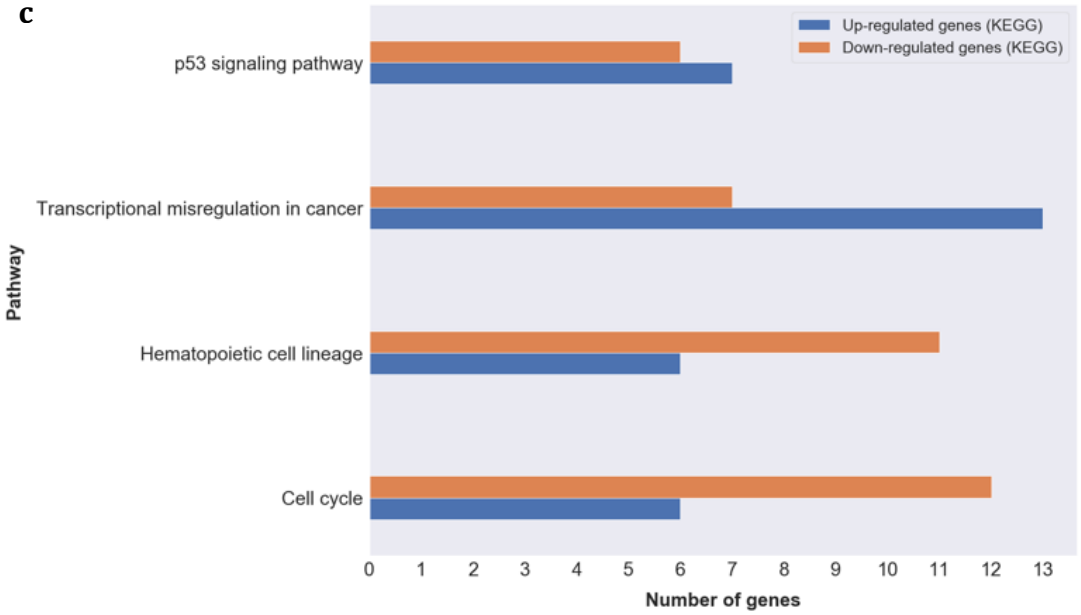


Figure 17 (cont'd)

d

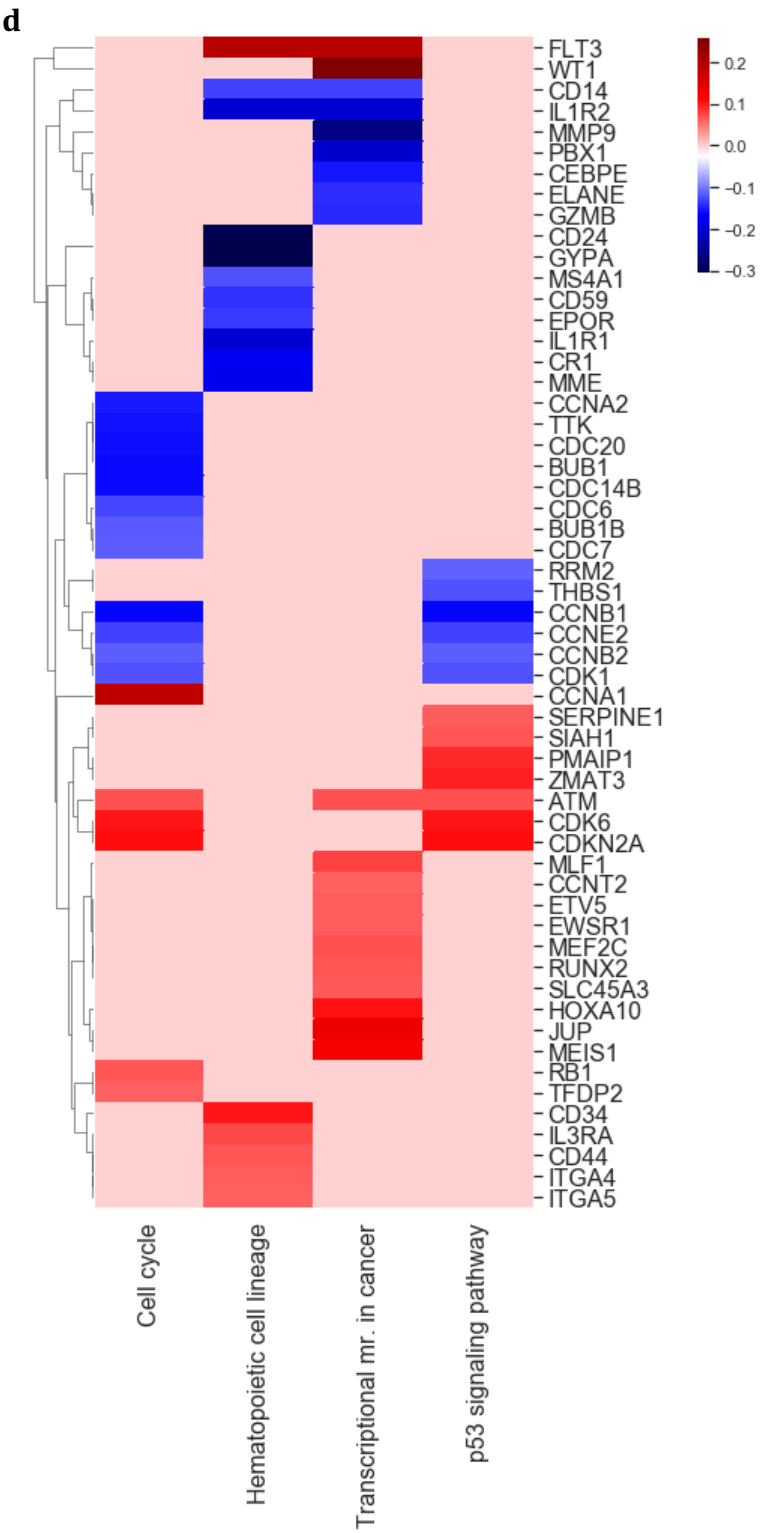


Figure 17 (cont'd)

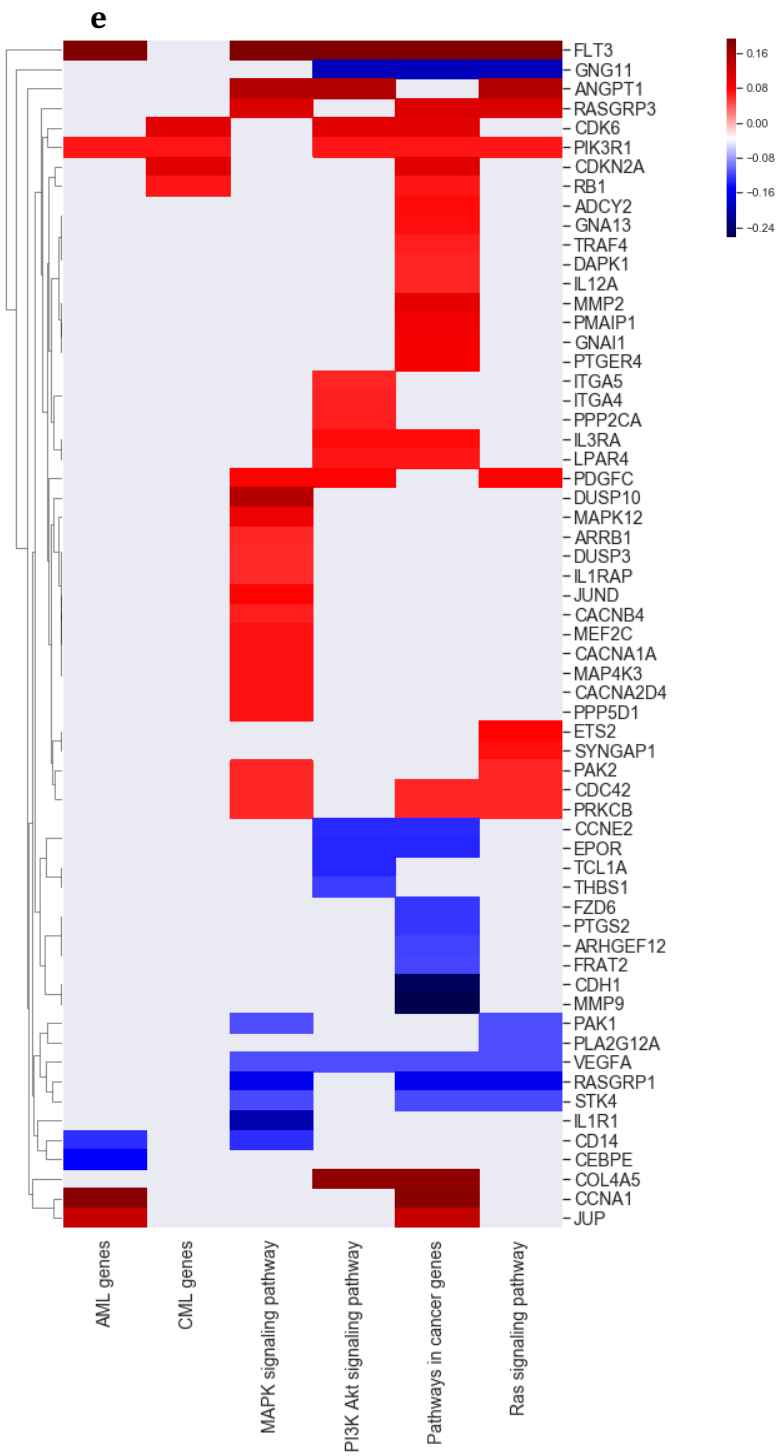
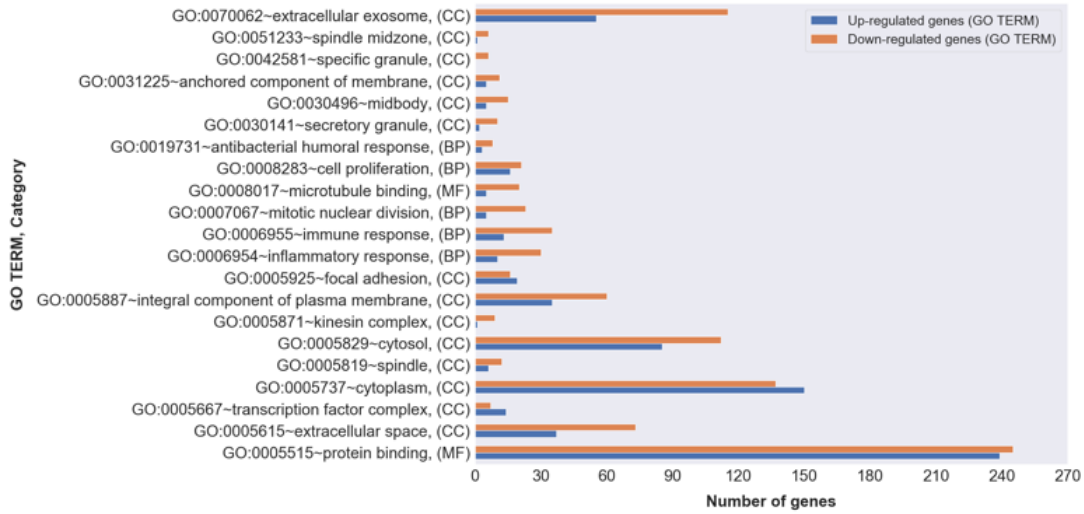


Figure 17 (cont'd)

f



For heatmaps, normalized values are represented in with blue for down-regulation and red for up-regulation, while light red/gray represents no reported specific direction. And for horizontal bar plot, same values are represented in with orange for down-regulation and blue for up-regulation**.** Heatmap of 974 DEPS (964 unique gene) in rows on 2,761 arrays (columns) including 2,213 AML patients and 548 healthy individuals from AML meta-analysis, using unsupervised hierarchical clustering and Euclidean distance for clustering. The age range of each age-groups is displayed in the legend and illustrated in the color bar on the top (labeled Age-group). The disease state (AML vs healthy) and sex of each subject are also represented in color bars on the top. **b.** Horizontal bar plot of the top 10 DEPS (gene symbols on vertical axis) from AML meta-analysis with mean difference values between AML and healthy (horizontal axis). **c.** Shows 4 KEGG signaling pathways deemed significant for our AML disease state enrichment analysis with number of up- and down-regulated DEPS enriched by each signaling pathway (horizontal axis), also visualized as a heatmap **(d)** of DEPS mean difference values with gene names (rows) identified in these 4 KEGG signaling pathways (columns). **e.** Shows

Figure 17 (cont'd)

the mean difference values with gene names (rows) of 61 DEPS enriched by 6 other KEGG signaling pathways (columns) pathways that are known to be involved in AML. Finally, the GO enrichment analysis results are summarized in **(f)**.

Figure 18. Sex-related gene expression meta-analysis in AML.

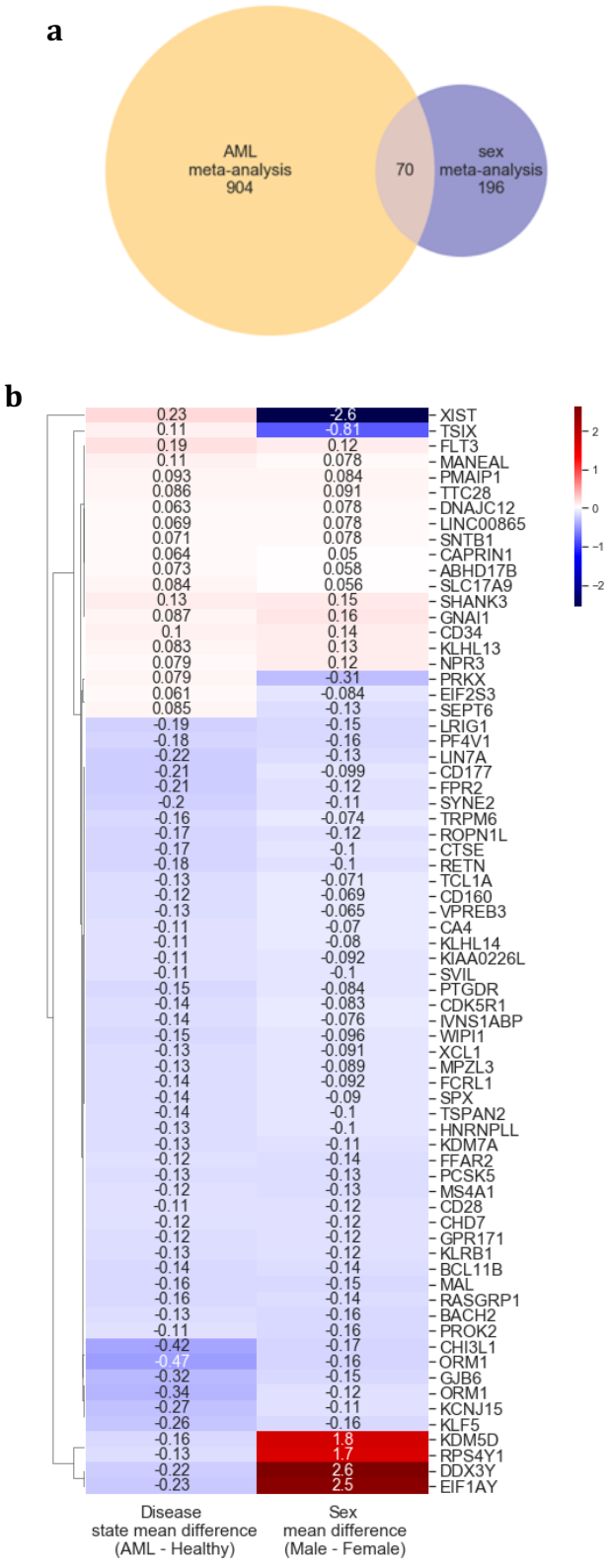


Figure 18 (cont'd)

c

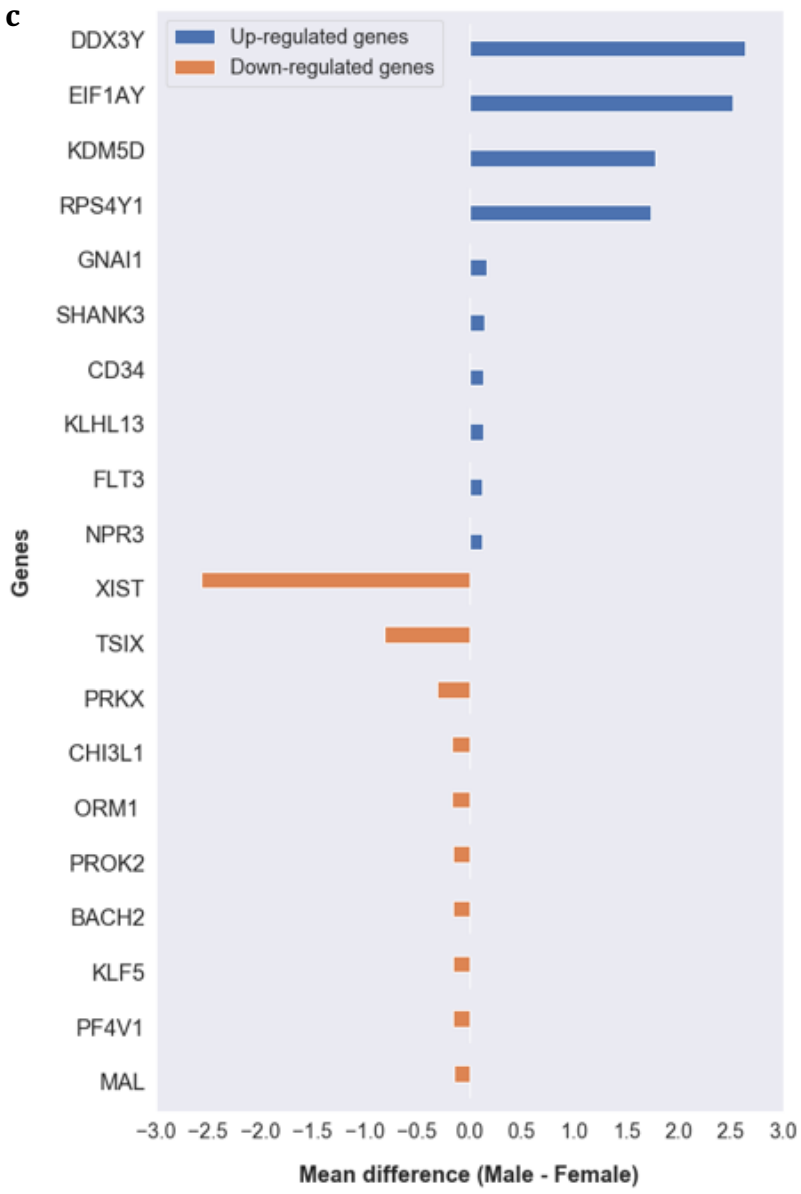


Figure 18 (cont'd)

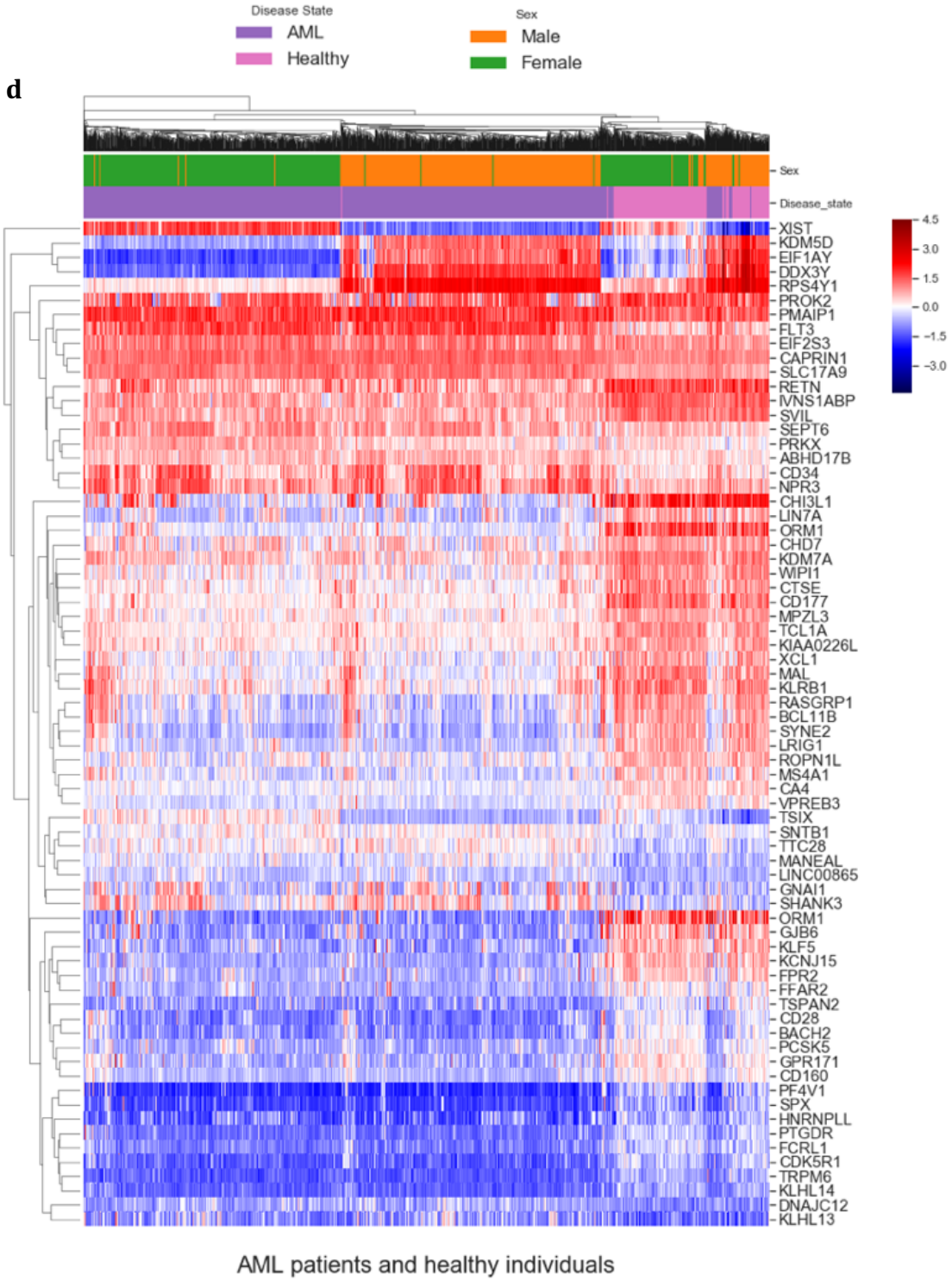
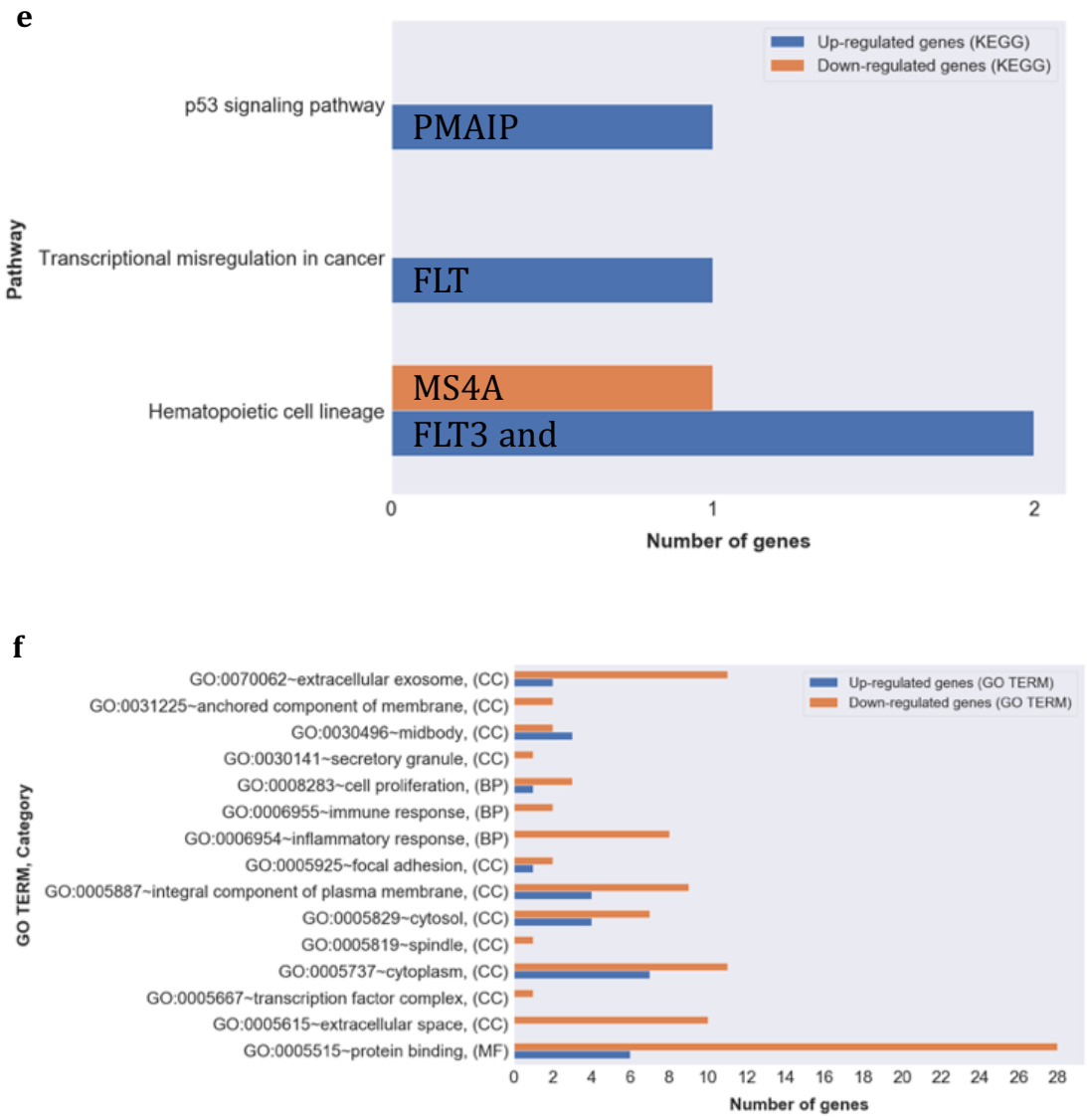


Figure 18 (cont'd)



a. the Venn diagram shows 70 DEPS identified (69 unique DE genes) to overlap with the DE genes from analysis 1, AML disease state meta-analysis. **b.** The heatmap of mean difference values comparison between the 70 DEPS overlapping genes between Analysis 1 and Analysis 2a. **c.** Horizontal bar plot of the top 10 DE genes from the 70 genes; genes are positioned at the y-axis, and x-axis represents mean difference values. **d.** Heatmap the 70 DEPS expression (rows) on 2,761 arrays (columns) including 2,213 AML patients and

Figure 18 (cont'd)

548 healthy individuals from Analysis 2a of sex-relevance in AML (using unsupervised hierarchical clustering and Euclidean distance for clustering). The disease state (AML vs healthy) and sex of each subject are indicated in color bars at the top. The disease state (AML vs healthy) and sex of each subject are indicated in color bars at the top. **e.** Pathway enrichment analysis using KEGG shows 3 signaling pathways were found enriched by 4 unique genes from this meta-analysis. **f.** Enrichment analysis for statistically significant overrepresented biological GO terms on the 70 DEPS genes.

Figure 19. Age-related gene expression meta-analysis in AML.

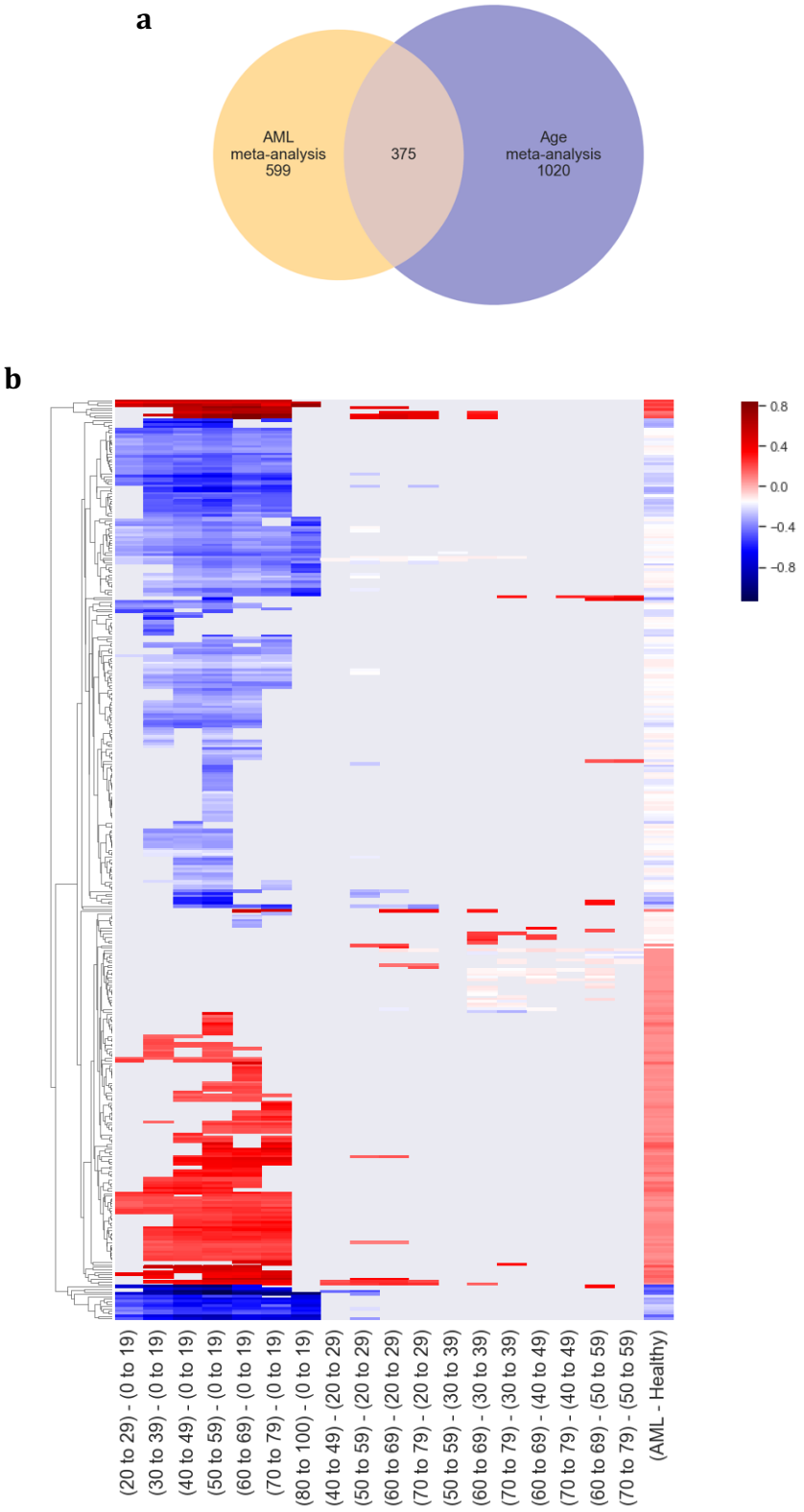


Figure 19 (cont'd)

c

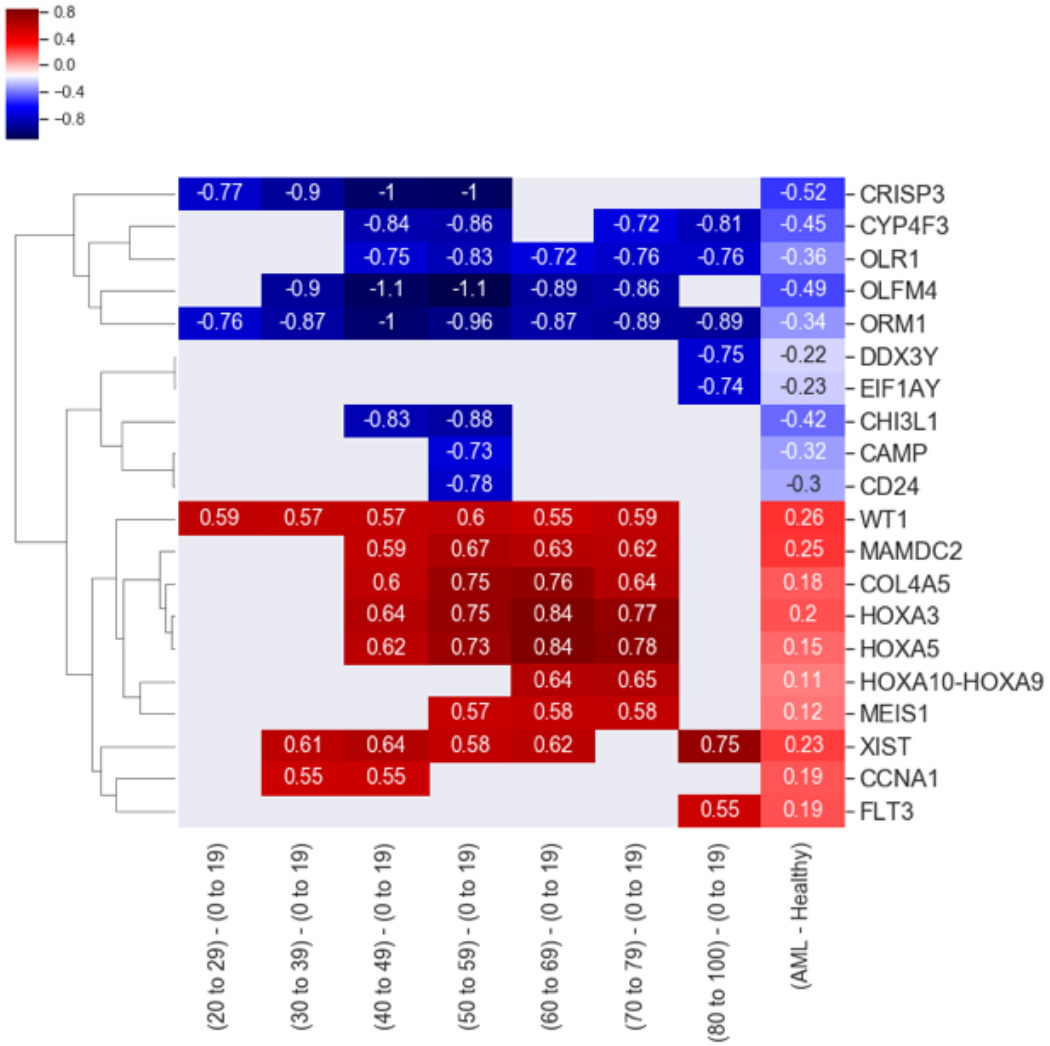


Figure 19 (cont'd)

d

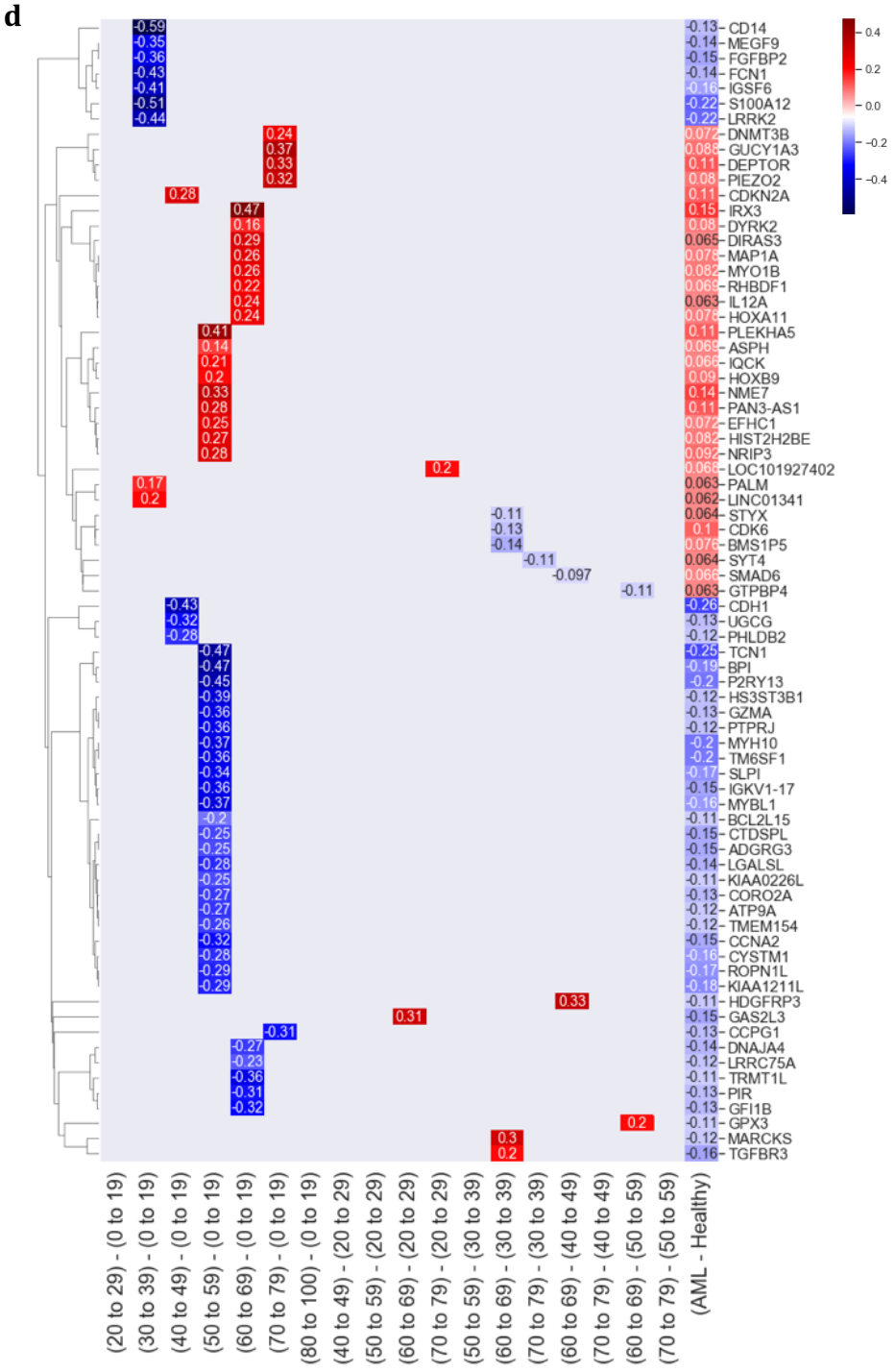


Figure 19 (cont'd)

e

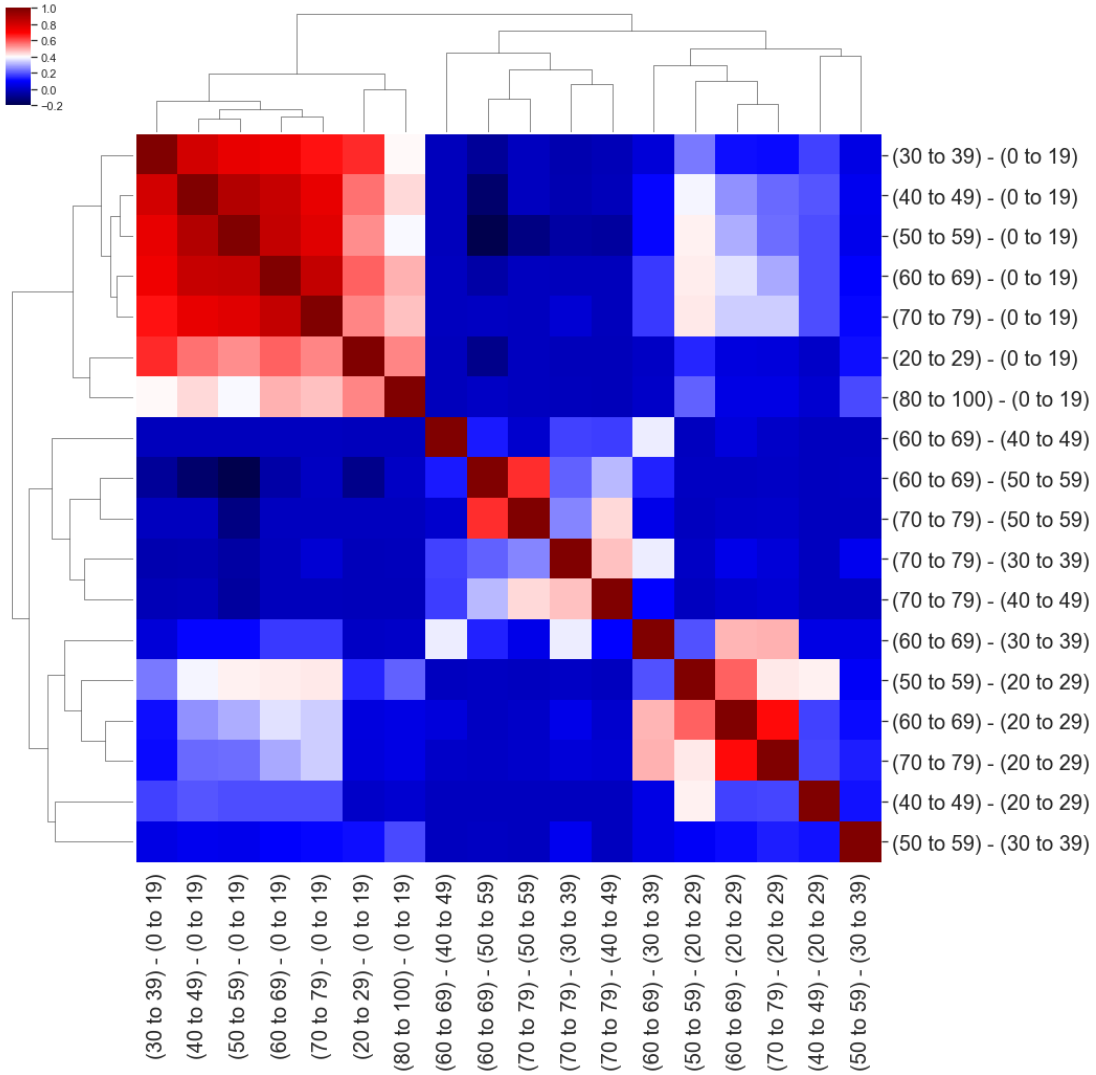


Figure 19 (cont'd)

f

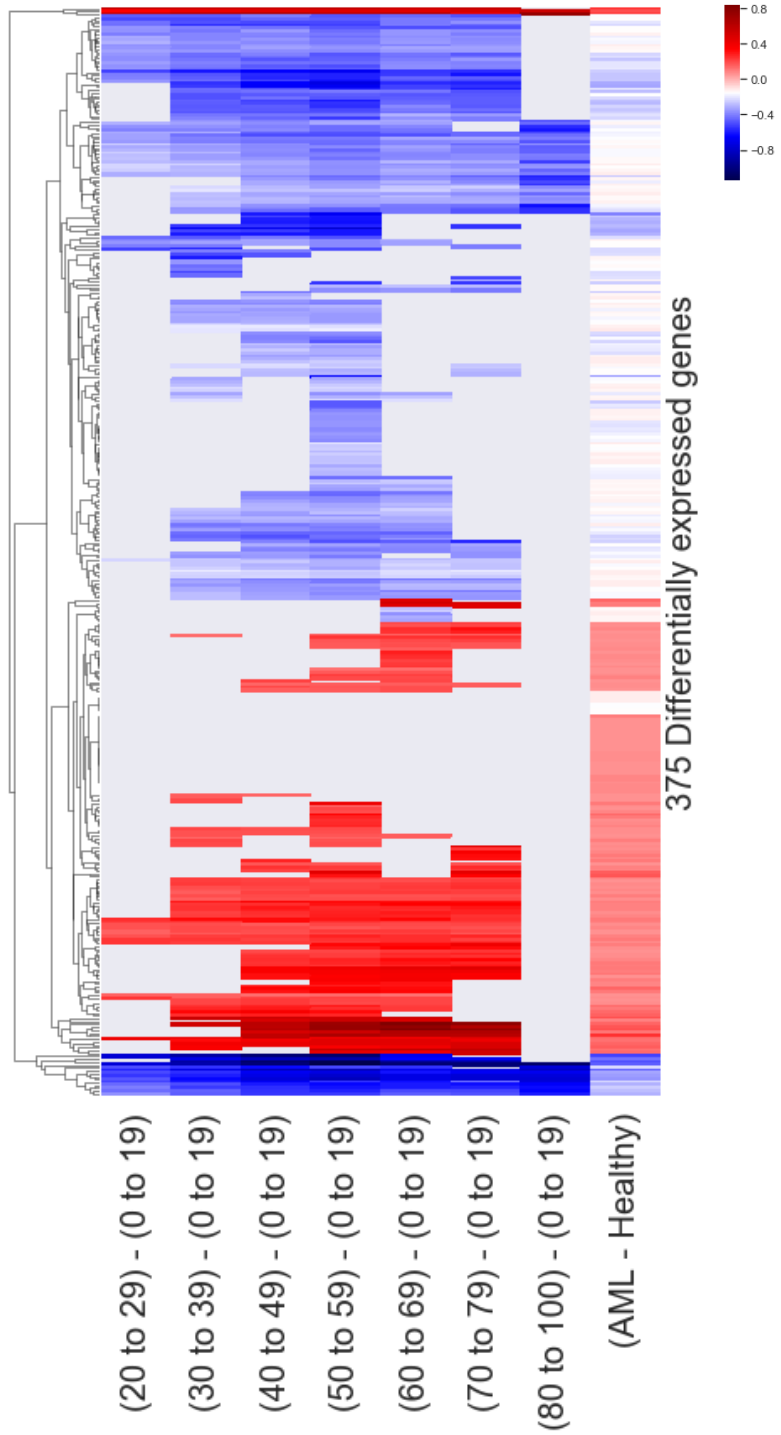


Figure 19 (cont'd)

g

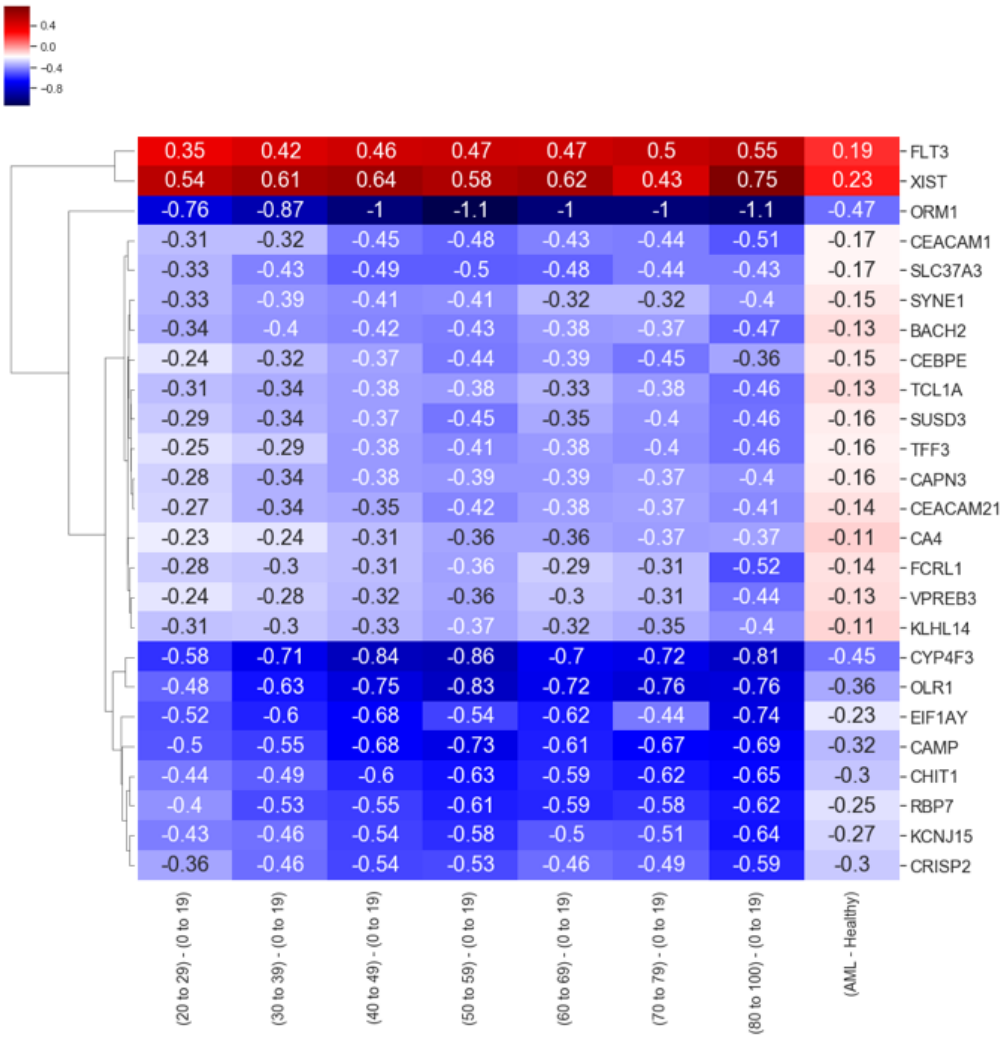
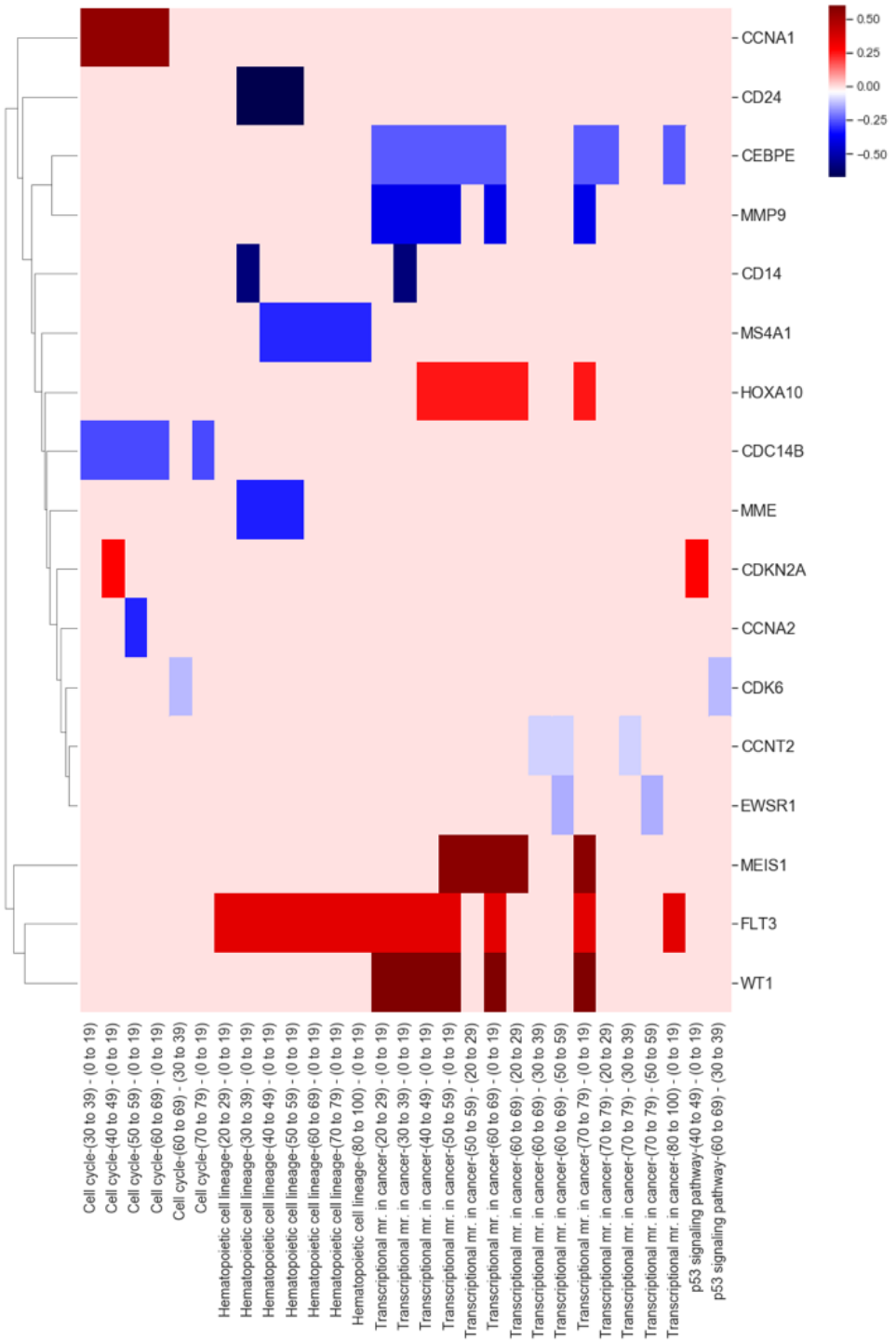


Figure 19 (cont'd)

h



For all heatmaps, normalized values are represented in with blue for down-regulation and red for up-regulation, while light red/gray represents no reported specific direction.

Figure 19. (cont'd)

Unsupervised hierarchical clustering and Euclidean distance was done for clustering on DEPS (vertical axis) mean difference values between each age-group comparison (horizontal axis). **a.** The Venn diagram shows 375 DEPS identified (372 unique DE genes) to overlap with the DE genes from analysis 1, AML disease state meta-analysis**. b.** Heatmap of 375 DEPS (rows) across 18 age-groups (columns) that were deemed statistically significant with mean difference values pf the 375 DEPS on vertical axis and age-groups on horizontal axis. **c.** Heatmap of the top 10 DE age-dependent genes (rows) mean difference values clustered on 7 age-groups (columns) with some genes appears in multiple age-groups while others appear only in one age-group. **d.** Shows 75 DEPS that are specific to a single age-group comparison. **e.** Age-group to age-group correlation matrix shows strong correlation direction between age-groups compared to the "0 to 19" age-group as a common reference. **f.** Shows the heatmap of 375 DEPS (rows) mean difference values across the "0 to 19" age-group (columns) as a common comparison reference for baseline analysis. **g.** Shows heatmap the mean difference values of 25 DE genes common across the baseline (0 to 19) age-group compared to 7 other age-groups that progress in age to illustrate gene expression changes with aging. We note that the mean difference values between AML and healthy cohorts from analysis 1 are shown in the right-most column of panels **(b-d), (f), (g)** for reference comparisons. **h.** Overlaps over KEGG pathways of 17 DE genes identified in 4 KEGG pathways according to age groups.

Table 5. Summary table of all 34 gene expression data sets used in our study.

Author, Year	GEO accession id	AML/Healthy		Affymetrix platform id: Number of samples used& Sample source	Refs.
Zatkova et al, 2009	GSE10258	AML		GPL570: 8 BM	[69]
Tomasson et al, 2008	GSE10358	AML		GPL570: 300 BM	[70]
Metzeler et al, 2008	GSE12417	AML		GPL570: 73 BM & 5 PB GPL96/97: 160 BM & 2PB	[54]
Wouters et al, 2009, Taskesen et al, 2011	GSE14468	AML		GPL570: 482 BM & 43 PB	[71,72]
Figueroa et al, 2009	GSE14479	AML		GPL570: 16 BM	[73]
Klein et al, 2009	GSE15434	AML		GPL570: 231 BM & 20 PB	[74]
Lück et al, 2011	GSE29883	AML		GPL570: 10 BM & 2 PB	[75]
Li et al, 2013, Herold et al, 2014, Janke et al, 2014, Jiang et al, 2016	GSE37642	AML		GPL570: 140 BM GPL96/97: 422 BM	[55-58]
Bullinger et al, 2014	GSE39363	AML		GPL570: 11 BM & 2 PB	NYP
Opel et al, 2015	GSE46819	AML		GPL570: 8 BM & 4 PB	[76]
TCGA et al, 2015	GSE68833	AML		GPL570: 183 BM	NYP
Cao et al, 2016	GSE69565	AML		GPL570: 12 PB	[77]
Bohl et al, 2016	GSE84334	AML		GPL570: 25 BM & 20 PB	NYP
Li et al, 2011	GSE23025	AML		GPL570: 21 BM & 13 PB	[78]
Warren et al, 2009	GSE11375	Healthy		GPL570: 26 PB	[79]
Green et al, 2009	GSE14845	Healthy		GPL570: 1 PB	NYP
Wu et al, 2012	GSE15932	Healthy		GPL570: 8 PB	NYP
Karlovich et al, 2009	GSE16028	Healthy		GPL570: 22 PB	[80]
Krug et al, 2011	GSE17114	Healthy		GPL570: 14 PB	NYP
Kong et al, 2012	GSE18123	Healthy		GPL570: 17 PB	[81]
Sharma et al, 2009	GSE18781	Healthy		GPL570: 25 PB	[82]
Rosell et al, 2011	GSE25414	Healthy		GPL570: 12 PB	[83]
Schmidt et al, 2006	GSE2842	Healthy		GPL570: 2 PB	[84]
Meng et al, 2015	GSE71226	Healthy		GPL570: 3 PB	NYP
Tasaki et al, 2017	GSE84844	Healthy		GPL570: 30 PB	[85]
Leday et al, 2018	GSE98793	Healthy		GPL570: 64 PB	[86]
Shamir et al, 2017	GSE99039	Healthy		GPL570: 121 PB	[87]
Tasaki et al, 2018	GSE93272	Healthy		GPL570: 35 PB	[68]
Clelland et al, 2013	GSE46449	Healthy		GPL570: 24 PB	[88]
Lauwerys et al, 2013 Ducreux et al, 2016	GSE39088	Healthy		GPL570: 46 PB	[89,90]
Xiao et al, 2011	GSE36809	Healthy		GPL570: 35 PB	[91]
Zhou et al, 2010	GSE19743	Healthy		GPL570: 63 PB	[92]
Jiang et al, 2018	GSE107968*	2 AML	1 Healthy	GPL570: 3 BM	NYP
Greiner et al, 2015	GSE68172*	20 AML	5 Healthy	GPL570: 25 PB	[64]
Majeti et al, 2009	GSE17054*	9 AML	4 Healthy	GPL570: 13 BM	[65]
Bacher et al, 2012	GSE33223*	20 AML	10 Healthy	GPL570: 30 PB	[66]
Mills et al, 2009	GSE15061*	404 AML	138 Healthy	GPL570: 542 BM	[67]

Meta-analysis data sets summary

Disease state		Sample source		Affymetrix platform id		Unique probesets	
AML	Healthy	BM	PB	GPL570	GPL96/97	GPL570	GPL96/97
2213	548	2090	671	2177	584	54,675	44,760

Table 5. (cont'd)

GEO, Gene Expression Omnibus; AML, acute myeloid leukemia; Ref. reference; NYP, not yet published, GPL570, Affymetrix Human Genome U133 Plus 2.0 Array; GPL96, Affymetrix Human Genome U133A Array; GPL97, Affymetrix Human Genome U133B Array; BM, Bone Marrow; PB, Peripheral Blood. A summary table of all our data sets using in our meta-analysis and disease classification. *"Covariate data sets," 5 data sets that were used during the batch correction step., data sets used only during the batch correction step to balance/account for batch in our curated data.

Table 6. Top 10 up- and down-regulated of DE genes in AML from disease state meta-analysis.

Up-regulated			
DEG name	**DEG Symbol**	**Tukey's HSD Mean difference**	**Bonferroni (p-adjusted)**
Wilms tumor 1	WT1	0.255353	< 4.11E-11
MAM domain containing 2	MAMDC2	0.248983	5.47E-09
X inactive specific transcript (non-protein coding)	XIST	0.230331	< 4.11E-11
homeobox A3	HOXA3	0.195790	1.1E-06
fms-related tyrosine kinase 3	FLT3	0.193420	< 4.11E-11
cyclin A1	CCNA1	0.185050	1.35E-07
mex-3 RNA binding family member B	MEX3B	0.181068	< 4.11E-11
collagen, type IV, alpha 5	COL4A5	0.177721	1.7E-05
neurexin 2	NRXN2	0.166598	< 4.11E-11
ATPase, Na+/K+ transporting, beta 1 polypeptide	ATP1B1	0.165197	5.47E-09
Down-regulated			
cysteine-rich secretory protein 3	CRISP3	-0.51965625	< 4.11E-11
olfactomedin 4	OLFM4	-0.489845396	< 4.11E-11
orosomucoid 1	ORM1	-0.465232864	< 4.11E-11
cytochrome P450, family 4, subfamily F, polypeptide 3	CYP4F3	-0.453467442	< 4.11E-11
chitinase 3-like 1 (cartilage glycoprotein-39)	CHI3L1	-0.421520435	< 4.11E-11
annexin A3	ANXA3	-0.390688999	< 4.11E-11
oxidized low density lipoprotein (lectin-like) receptor 1	OLR1	-0.35525472	< 4.11E-11
carcinoembryonic antigen-related cell adhesion molecule 8	CEACAM8	-0.351181264	< 4.11E-11
orosomucoid 1	ORM1	-0.336303304	< 4.11E-11
tumor-associated calcium signal transducer 2	TACSTD2	-0.323939961	< 4.11E-11

From the Post-hoc Tukey's test, gene expression means difference value < 5% or > 95% between AML and healthy (AML - healthy) were deemed statistically significant for AML. Genes were considered disease state statistically significant from the analysis of all 2761 cases (2213 AML patients and 548 healthy controls) using. The p-values were adjusted based on Bonferroni correction for false discovery rate (FDR). Significant DE genes are listed in descending order of the mean difference value comparisons for disease state.

Table 7. KEGG functional analysis of 974 DEPS from meta-analysis of 34 gene expression data sets.

Pathway	No. of genes	Down-regulated	Up-regulated	p-value (unadjusted)	Benjamini (p-adjusted)
Hematopoietic cell lineage	11, 6	IL1R2, CD59, GYPA, MS4A1, EPOR, CD24, CD14, EPOR, IL1R1, MME, CR1	ITGA4, FLT3, CD34, IL3RA, ITGA5, CD44	2.3E-5	5.8E-3
Cell cycle	12, 6	CDC7, CDC6, CCNB1, CDC20, CCNA2, CCNE2, TTK, CDC14B, CDK1, BUB1, CCNB2, BUB1B	6 RB1, CCNA1, CDK6, ATM, TFDP2, CDKN2A	1.4E-4	1.2E-2
p53 signaling pathway	6, 7	THBS1, CCNB1, CCNE2, CDK1, RRM2, CCNB2	SIAH1, CDK6, ATM, SERPINE1, CDKN2A, PMAIP1, ZMAT3	1.0E-4	1.3E-2
Transcriptional misregulation in cancer	7, 13	IL1R2, GZMB, CD14, ELANE, MMP9, CEBPE, PBX1	WT1, RUNX2, ETV5, MEIS1, JUP, EWSR1, ATM, HOXA10, MLF1, FLT3, CCNT2, MEF2C, SLC45A3	6.5E-4	4.1E-2

974 DEPS (487 overexpressed and 487 underexpressed) from analysis 1 were enriched by 4 statistically significant KEGG pathways. Signaling pathways were deemed

Table 7. (cont'd)

significant based on Benjamini- Hochberg adjusted p-value < 0.05. The two numbers in each cell in "No. of genes" column indicate the down-regulated **(first)** and up-regulated **(second)** DE genes that enriched by each pathway. Pathways are listed in descending order using of Benjamini- Hochberg adjusted p-value.

Table 8. AML sex relevance (male - female) DE genes & associated signaling pathways.

Pathway	No. of genes	High in Females	High in Males
Hematopoietic cell lineage	1, 2	MS4A1	FLT3, CD34
p53 signaling pathway	–, 1	–	PMAIP1
Transcriptional misregulation in cancer	–, 1	–	FLT3

Common DEPS between the 70 DEPS from analysis 2a and the 974 DEPS from AML disease state meta-analysis for KEGG pathways and GO terms. 4 sex-relevant unique DE genes in AML were found in 3 different signaling pathways, including, 1 highly expressed in females and 3 highly expressed in males.

Table 9. AML age-dependent (AML - healthy) DE genes & associated signaling pathways.

Pathway	No. of genes	Down-regulated Age-group	Up-regulated Age-group
Hematopoietic cell lineage	4, 1	CD14 (30 to 39) - (0 to 19)	FLT3 (20 to 29) - (0 to 19), (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (60 to 69) - (0 to 19), (70 to 79) - (0 to 19), (80 to 100) - (0 to 19)
		MME (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19)	
		CD24 (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19)	
		MS4A1 (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (60 to 69) - (0 to 19), (70 to 79) - (0 to 19), (80 to 100) - (0 to 19)	
Cell cycle	3, 2	CCNA2 (50 to 59) - (0 to 19)	CCNA1 (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (60 to 69) - (0 to 19)
		CDK6 (60 to 69) - (30 to 39)	
		CDC14B (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (60 to 69) - (0 to 19), (70 to 79) - (0 to 19)	CDKN2A (40 to 49) - (0 to 19)
p53 signaling pathway	1, 1	CDK6 (60 to 69) - (30 to 39)	CDKN2A (40 to 49) - (0 to 19)
Transcriptional misregulation in cancer	5, 4	CD14 (30 to 39) - (0 to 19)	MEIS1 (50 to 59) - (0 to 19), (50 to 59) - (20 to 29), (60 to 69) - (0 to 19), (60 to 69) - (20 to 29), (70 to 79) - (0 to 19)
		MMP9 (20 to 29) - (0 to 19), (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (60 to 69) - (0 to 19), (70 to 79) - (0 to 19)	
		EWSR1 (60 to 69) - (50 to 59), (70 to 79) - (50 to 59)	WT1 (20 to 29) - (0 to 19), (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (60 to 69) - (0 to 19), (70 to 79) - (0 to 19)
		CEBPE (20 to 29) - (0 to 19), (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (50 to 59) - (20 to 29), (60 to 69) - (0 to19), (70 to 79) - (0 to 19), (70 to 79) - (20 to29), (80 to 100) - (0 to 19)	FLT3 (20 to 29) - (0 to 19), (30 to 39) - (0 to 19), (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (60 to 69) - (0 to 19), (70 to 79) - (0 to 19), (80 to 100) - (0 to 19)
		CCNT2 (60 to 69) - (30 to 39), (70 to 79) - (30 to 39), (60 to 69) - (50 to 59)	HOXA10 (40 to 49) - (0 to 19), (50 to 59) - (0 to 19), (50 to 59) - (20 to 29), (60 to 69) - (0 to 19), (60 to 69) - (20 to 29), (70 to 79) - (0 to 19)

Table 9. (cont'd)

Common DE genes between the 375 DEPS from analysis 2b and the 974 DEPS from analysis 1 for KEGG pathways and GO terms. 17 age-dependent unique DE genes in AML were found in 4 different signaling pathways. DE genes are listed according to associated age-groups for each signaling pathway.

Table 10. Age-dependent genes show drug to gene interaction.

Druggable Gene Category	Matching gene count	Matching genes(s) DGIdb analysis 1	Matching genes(s) DGIdb analysis 2
DRUGGABLE GENOME	15, 24	TFF3, ORM1, CA4, CYP4F2, CYP4F3, CEACAM1, FLT3, CHIT1, OLR1, KCNJ15, CAMP, CRISP2, CAPN3, SLC37A3, FCRL1	CDH1, GPX3, CD14, DYRK2, SLPI, CCNA2, TGFBR3, UGCG, FCN1, GZMA, TCN1, BPI, S100A12, CDK6, IL12A, P2RY13, ADGRG3, DNMT3B, GUCY1A3, FGFBP2, PTPRJ, LRRK2, BCL2L15, STYX
KINASE	3, 12	CEACAM1, FLT3, TCL1A	DYRK2, CCNA2, TGFBR3, S100A12, CDKN2A, CDK6, DIRAS3, GTPBP4, DEPTOR, PTPRJ, NME7, LRRK2
SERINE THREONINE KINASE	2, 10	FLT3, TCL1A	DYRK2, CCNA2, TGFBR3, S100A12, CDKN2A, CDK6, DIRAS3, GTPBP4, PTPRJ, LRRK2
TUMOR SUPPRESSOR	–, 8	–	CTDSPL, CCNA2, CDKN2A, CDK6, IL12A, DIRAS3, GTPBP4, CCPG1
CELL SURFACE	2, 5	CA4, CEACAM1	CD14, TGFBR3, FCN1, MYH10, PTPRJ
PROTEASE	–, 6	–	SLPI, FCN1, GZMA, ASPH, IGKV1-17, NRIP3
CLINICALLY ACTIONABLE	1, 4	FLT3	CDH1, CDKN2A, CDK6, DNMT3B
TRANSPORTER	4, –	CEACAM1, KCNJ15, SLC37A3, RBP7	–
TRANSCRIPTION FACTOR COMPLEX	–, 4	–	SMAD6, GFI1B, HOXA11, HOXB9
EXTERNAL SIDE OF PLASMA MEMBRANE	1, 3	CA4	CD14, TGFBR3, FCN1
HISTONE MODIFICATION	–, 3	–	CCNA2, GFI1B, DNMT3B
CYTOCHROME P450	2, –	CYP4F2, CYP4F3	–
DRUG METABOLISM	–, 1	CYP4F2	–
EXCHANGER	–, 1	SLC37A3	–

Two analysis were carried using DGIdb, (i) 25 DE genes common across the baseline (0 to 19) age-group and (ii) 75 genes identified to be specific to one age-group were used to carry out drug-gene interaction analysis using DGIdb. Functional classes for both

Table 10. (cont'd)

analyses are shown here. "Matching gene count" column, which indicates the matching genes between from analysis (i) **(first number)** and matching genes from DGIdb analysis (ii) **(second number)** DE genes that enriched by each pathway.

BIBLIOGRAPHY

BIBLIOGRAPHY

1 Kumar, C. C. Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes Cancer* **2**, 95-107, doi:10.1177/1947601911408076 (2011).

2 De Kouchkovsky, I. & Abdul-Hay, M. 'Acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood Cancer J* **6**, e441, doi:10.1038/bcj.2016.50 (2016).

3 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7-30, doi:10.3322/caac.21442 (2018).

4 Institute, N. C. SEER Cancer Stat Facts: Acute Myeloid Leukemia (Percent of New Cases by Age Group). [https://seer.cancer.gov/statfacts/html/amyl.html]. ((accessed 11.30.18), 2011-2015).

5 Short, N. J., Rytting, M. E. & Cortes, J. E. Acute myeloid leukaemia. *Lancet* **392**, 593-606, doi:10.1016/S0140-6736(18)31041-9 (2018).

6 Dohner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N Engl J Med* **373**, 1136-1152, doi:10.1056/NEJMra1406184 (2015).

7 Reese, N. D. & Schiller, G. J. High-dose cytarabine (HD araC) in the treatment of leukemias: a review. *Curr Hematol Malig Rep* **8**, 141-148, doi:10.1007/s11899-013-0156-3 (2013).

8 Dohner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453-474, doi:10.1182/blood-2009-07-235358 (2010).

9 Meyers, J., Yu, Y., Kaye, J. A. & Davis, K. L. Medicare fee-for-service enrollees with primary acute myeloid leukemia: an analysis of treatment patterns, survival, and healthcare resource utilization and costs. *Appl Health Econ Health Policy* **11**, 275-286, doi:10.1007/s40258-013-0032-2 (2013).

10 Ferrara, F. & Schiffer, C. A. Acute myeloid leukaemia in adults. *Lancet* **381**, 484-495, doi:10.1016/S0140-6736(12)61727-9 (2013).

11 Appelbaum, F. R. *et al.* Age and acute myeloid leukemia. *Blood* **107**, 3481-3485, doi:10.1182/blood-2005-09-3724 (2006).

12 Grimwade, D. & Hills, R. K. Independent prognostic factors for AML outcome. *Hematology Am Soc Hematol Educ Program*, 385-395, doi:10.1182/asheducation-2009.1.385 (2009).

13 Dohner, H. Implication of the molecular characterization of acute myeloid leukemia. *Hematology Am Soc Hematol Educ Program*, 412-419, doi:10.1182/asheducation-2007.1.412 (2007).

14 Walter, M. J. *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci U S A* **106**, 12950-12955, doi:10.1073/pnas.0903091106 (2009).

15 Suela, J., Alvarez, S. & Cigudosa, J. C. DNA profiling by arrayCGH in acute myeloid leukemia and myelodysplastic syndromes. *Cytogenet Genome Res* **118**, 304-309, doi:10.1159/000108314 (2007).

16 Armstrong, S. A. *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics* **30**, 41-47, doi:10.1038/ng765 (2002).

17 Debernardi, S. *et al.* Genome-wide analysis of acute myeloid leukemia with normal karyotype reveals a unique pattern of homeobox gene expression distinct from those with translocation-mediated fusion events. *Gene Chromosome Canc* **37**, 149-158, doi:10.1002/gcc.10198 (2003).

18 Schoch, C. *et al.* Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. *P Natl Acad Sci USA* **99**, 10008-10013, doi:10.1073/pnas.142103599 (2002).

19 Miller, B. G. & Stamatoyannopoulos, J. A. Integrative meta-analysis of differential gene expression in acute myeloid leukemia. *PLoS One* **5**, e9466, doi:10.1371/journal.pone.0009466 (2010).

20 Ramasamy, A., Mondry, A., Holmes, C. C. & Altman, D. G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* **5**, e184, doi:10.1371/journal.pmed.0050184 (2008).

21 Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127, doi:10.1093/biostatistics/kxj037 (2007).

22 Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* **6**, e17238, doi:10.1371/journal.pone.0017238 (2011).

23 Pavlidis, P. Using ANOVA for gene selection from microarray studies of the nervous system. *Methods* **31**, 282-289, doi:10.1016/S1046-2023(03)00157-9 (2003).

24 Pavlidis, P. & Noble, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **19**, 295-296, doi:DOI 10.1093/bioinformatics/19.2.295 (2003).

25 Mias, G. in *Mathematica for Bioinformatics: A Wolfram Language Approach to Omics* 193-226 (Springer International Publishing, 2018).

26 Neyman, J. & Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* **20a**, 263-294, doi:DOI 10.1093/biomet/20A.3-4.263 (1928).

27 Waltman, L. & Schreiber, M. On the calculation of percentile-based bibliometric indicators. *J Am Soc Inf Sci Tec* **64**, 372-379, doi:10.1002/asi.22775 (2013).

28 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361, doi:10.1093/nar/gkw1092 (2017).

29 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457-D462, doi:10.1093/nar/gkv1070 (2016).

30 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

31 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

32 Carbon, S. *et al.* Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45**, D331-D338, doi:10.1093/nar/gkw1108 (2017).

33 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).

34 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

35 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).

36 Griffith, M. *et al.* DGIdb: mining the druggable genome. *Nat Methods* **10**, 1209-+, doi:10.1038/Nmeth.2689 (2013).

37 Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391-2405, doi:10.1182/blood-2016-03-643544 (2016).

38 Martelli, M. P., Sportoletti, P., Tiacci, E., Martelli, M. F. & Falini, B. Mutational landscape of AML with normal cytogenetics: biological and clinical implications. *Blood Rev* **27**, 13-22, doi:10.1016/j.blre.2012.11.001 (2013).

39 Klepin, H. D., Rao, A. V. & Pardee, T. S. Acute myeloid leukemia and myelodysplastic syndromes in older adults. *J Clin Oncol* **32**, 2541-2552, doi:10.1200/JCO.2014.55.1564 (2014).

40 Cancer Genome Atlas Research, N. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).

41 Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* **366**, 1090-1098, doi:10.1056/NEJMoa1106968 (2012).

42 Hou, H. A. *et al.* WT1 mutation in 470 adult patients with acute myeloid leukemia: stability during disease evolution and implication of its incorporation into a survival scoring system. *Blood* **115**, 5222-5231, doi:10.1182/blood-2009-12-259390 (2010).

43 Ho, P. A. *et al.* Prevalence and prognostic implications of WT1 mutations in pediatric acute myeloid leukemia (AML): a report from the Children's Oncology Group. *Blood* **116**, 702-710, doi:10.1182/blood-2010-02-268953 (2010).

44 Udby, L., Calafat, J., Sorensen, O. E., Borregaard, N. & Kjeldsen, L. Identification of human cysteine-rich secretory protein 3 (CRISP-3) as a matrix protein in a subset of peroxidase-negative granules of neutrophils and in the granules of eosinophils. *J Leukocyte Biol* **72**, 462-469 (2002).

45 Izzi, V. *et al.* An extracellular matrix signature in leukemia precursor cells and acute myeloid leukemia. *Haematologica* **102**, E245-E248, doi:10.3324/haematol.2017.167304 (2017).

46 Buggins, A. G. *et al.* Microenvironment produced by acute myeloid leukemia cells prevents T cell activation and proliferation by inhibition of NF-kappaB, c-Myc, and pRb pathways. *J Immunol* **167**, 6021-6030 (2001).

47 Rashidi, A. & Uy, G. L. Targeting the Microenvironment in Acute Myeloid Leukemia. *Curr Hematol Malig R* **10**, 126-131, doi:10.1007/s11899-015-0255-4 (2015).

48 Borrow, J. *et al.* The t(7;11)(p15;p15) translocation in acute myeloid leukaemia fuses the genes for nucleoporin NUP98 and class I homeoprotein HOXA9. *Nature Genetics* **12**, 159-167, doi:DOI 10.1038/ng0296-159 (1996).

49 Andreeff, M. *et al.* HOX expression patterns identify a common signature for favorable AML. *Leukemia* **22**, 2041-2047, doi:10.1038/leu.2008.198 (2008).

50 Fan, C., Stendahl, U., Stjernberg, N. & Beckman, L. Association between Orosomucoid Types and Cancer. *Oncology* **52**, 498-500 (1995).

51 Abbate, F., Casini, A., Owa, T., Scozzafava, A. & Supuran, C. T. Carbonic anhydrase inhibitors: E7070, a sulfonamide anticancer agent, potently inhibits cytosolic isozymes I and II, and transmembrane, tumor-associated isozyme IX. *Bioorg Med Chem Lett* **14**, 217-223, doi:10.1016/j.bmcl.2003.09.062 (2004).

52 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).

53 Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).

54 Metzeler, K. H. *et al.* An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* **112**, 4193-4201, doi:10.1182/blood-2008-02-134411 (2008).

55 Li, Z. *et al.* Identification of a 24-gene prognostic signature that improves the European LeukemiaNet risk classification of acute myeloid leukemia: an international collaborative study. *J Clin Oncol* **31**, 1172-1181, doi:10.1200/JCO.2012.44.3184 (2013).

56 Herold, T. *et al.* Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood* **124**, 1304-1311, doi:10.1182/blood-2013-12-540716 (2014).

57 Janke, H. *et al.* Activating FLT3 Mutants Show Distinct Gain-of-Function Phenotypes In Vitro and a Characteristic Signaling Pathway Profile Associated with Prognosis in Acute Myeloid Leukemia. *Plos One* **9**, doi:ARTN e89560 10.1371/journal.pone.0089560 (2014).

58 Jiang, X. *et al.* Eradication of Acute Myeloid Leukemia with FLT3 Ligand-Targeted miR-150 Nanoparticles. *Cancer Res* **76**, 4470-4480, doi:10.1158/0008-5472.CAN-15-2949 (2016).

59 Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

60 Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).

61 Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264, doi:10.1093/biostatistics/4.2.249 (2003).

62 Roushangar, R. & Mias, G. I. ClassificaIO: machine learning for classification graphical user interface. *bioRxiv*, doi:10.1101/240184 (2017).

63 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).

64 Schneider, V. Z., L.; Markus, R.; Fekete, N.; Schrezenmeier, H.; Erle, A. ; Lars, B.; Hofmann, S.; Götz, M.; Döhner, K.; Ihme, S.; Döhner, H.; Buske, C.; Feuring-Buske, M.; Greiner, J. Leukemic progenitor cells are susceptible to targeting by stimulated cytotoxic T cells against immunogenic leukemia-associated antigens. (2015).

65 Majeti, R. *et al.* Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc Natl Acad Sci U S A* **106**, 3396-3401, doi:10.1073/pnas.0900089106 (2009).

66 Bacher, U. *et al.* Multilineage dysplasia does not influence prognosis in CEBPA-mutated AML, supporting the WHO proposal to classify these patients as a unique entity. *Blood* **119**, 4719-4722, doi:10.1182/blood-2011-12-395574 (2012).

67 Mills, K. I. *et al.* Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood* **114**, 1063-1072, doi:10.1182/blood-2008-10-187203 (2009).

68 Tasaki, S. *et al.* Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat Commun* **9**, 2755, doi:10.1038/s41467-018-05044-4 (2018).

69 Zatkova, A. *et al.* AML/MDS with 11q/MLL amplification show characteristic gene expression signature and interplay of DNA copy number changes. *Genes Chromosomes Cancer* **48**, 510-520, doi:10.1002/gcc.20658 (2009).

70 Tomasson, M. H. *et al.* Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. *Blood* **111**, 4797-4808, doi:10.1182/blood-2007-09-113027 (2008).

71 Taskesen, E. *et al.* Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood* **117**, 2469-2475, doi:10.1182/blood-2010-09-307280 (2011).

72 Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088-3091, doi:10.1182/blood-2008-09-179895 (2009).

73 Figueroa, M. E. *et al.* Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood* **113**, 2795-2804, doi:10.1182/blood-2008-08-172387 (2009).

74 Klein, H. U. *et al.* Quantitative comparison of microarray experiments with published leukemia related gene expression signatures. *BMC Bioinformatics* **10**, 422, doi:10.1186/1471-2105-10-422 (2009).

75 Luck, S. C. *et al.* Deregulated apoptosis signaling in core-binding factor leukemia differentiates clinically relevant, molecular marker-independent subgroups. *Leukemia* **25**, 1728-1738, doi:10.1038/leu.2011.154 (2011).

76 Opel, D. *et al.* Targeting inhibitor of apoptosis proteins by Smac mimetic elicits cell death in poor prognostic subgroups of chronic lymphocytic leukemia. *Int J Cancer* **137**, 2959-2970, doi:10.1002/ijc.29650 (2015).

77 Cao, Q. *et al.* BCOR regulates myeloid cell proliferation and differentiation. *Leukemia* **30**, 1155-1165, doi:10.1038/leu.2016.2 (2016).

78 Li, L. *et al.* Altered hematopoietic cell gene expression precedes development of therapy-related myelodysplasia/acute myeloid leukemia and identifies patients at risk. *Cancer Cell* **20**, 591-605, doi:10.1016/j.ccr.2011.09.011 (2011).

79 Warren, H. S. *et al.* A genomic score prognostic of outcome in trauma patients. *Mol Med* **15**, 220-227, doi:10.2119/molmed.2009.00027 (2009).

80 Karlovich, C. *et al.* A longitudinal study of gene expression in healthy individuals. *BMC Med Genomics* **2**, 33, doi:10.1186/1755-8794-2-33 (2009).

81 Kong, S. W. *et al.* Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One* **7**, e49475, doi:10.1371/journal.pone.0049475 (2012).

82 Sharma, S. M. *et al.* Insights in to the pathogenesis of axial spondyloarthropathy based on gene expression profiles. *Arthritis Res Ther* **11**, R168, doi:10.1186/ar2855 (2009).

83 Rosell, A. *et al.* Brain perihematoma genomic profile following spontaneous human intracerebral hemorrhage. *PLoS One* **6**, e16750, doi:10.1371/journal.pone.0016750 (2011).

84 Schmidt, S. *et al.* Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia. *Blood* **107**, 2061-2069, doi:10.1182/blood-2005-07-2853 (2006).

85 Tasaki, S. *et al.* Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjogren's syndrome. *Ann Rheum Dis* **76**, 1458-1466, doi:10.1136/annrheumdis-2016-210788 (2017).

86 Leday, G. G. R. *et al.* Replicable and Coupled Changes in Innate and Adaptive Immune Gene Expression in Two Case-Control Studies of Blood Microarrays in Major Depressive Disorder. *Biol Psychiatry* **83**, 70-80, doi:10.1016/j.biopsych.2017.01.021 (2018).

87 Shamir, R. *et al.* Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **89**, 1676-1683, doi:10.1212/WNL.0000000000004516 (2017).

88 Clelland, C. L. *et al.* Utilization of never-medicated bipolar disorder patients towards development and validation of a peripheral biomarker profile. *PLoS One* **8**, e69082, doi:10.1371/journal.pone.0069082 (2013).

89 Ducreux, J. *et al.* Interferon alpha kinoid induces neutralizing anti-interferon alpha antibodies that decrease the expression of interferon-induced and B cell activation associated transcripts: analysis of extended follow-up data from the interferon alpha kinoid phase I/II study. *Rheumatology (Oxford)* **55**, 1901-1905, doi:10.1093/rheumatology/kew262 (2016).

90 Lauwerys, B. R. *et al.* Down-regulation of interferon signature in systemic lupus erythematosus patients by active immunization with interferon alpha-kinoid. *Arthritis Rheum* **65**, 447-456, doi:10.1002/art.37785 (2013).

91 Xiao, W. *et al.* A genomic storm in critically injured humans. *J Exp Med* **208**, 2581-2590, doi:10.1084/jem.20111354 (2011).

92 Zhou, B. *et al.* Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proc Natl Acad Sci U S A* **107**, 9923-9928, doi:10.1073/pnas.1002757107 (2010).

Chapter 4 –

Summary and Outlook

Conclusion

In this dissertation, we aimed to establish sex-related and age-dependent DE genes with related gene expression patterns and associated signaling pathways as biomarkers in AML. Our approach utilized machine learning methods, which led to the development of a graphical user interface to facilitate model training and testing for classification, (Chapter 2). Subsequently, we carried 3 gene expression meta-analyses and gene enrichment analyses on publicly available gene expression data, accumulated from 2,761 subjects (2,213 AML patients and 548 healthy individuals). We analyzed a total of 44,754 probe sets (corresponding to multiple genes) per subject. We used multiple statistical methods for microarray analysis were used to pre-process raw data, and also implemented a "data-wise" batch effect correction. The latter was used to correct for batch effects caused by study variability and sample processing. Following normalization and batch effect correction across arrays, we used a statistical linear model to study the effects of age and sex on gene expression in AML patients as compared to healthy individuals (Chapter 3). Three downstream differential gene expression analyses were carried out:

Analysis 1: Gene expression meta-analysis and associated signaling pathways of AML disease state compared to healthy individuals. From this analysis we identified 964 DE unique genes (974 DEPS) including 56 DE unique genes (27 up- and 29 down-regulated) that were associated with 4 statistically significant KEGG pathways including Hematopoietic cell lineage, Cell cycle, p53 signaling pathway, and Transcriptional misregulation in cancer. Multiple genes identified do not have known associations with

AML and signaling pathways and can provide new avenues of investigation and novel hypothesis-driven mechanistic studies.

Analysis 2a: Sex-dependent gene expression meta-analysis and associated signaling pathways in AML compared to healthy individuals, from this analysis we identified 70 DEPS with 69 unique DE genes that overlapped between analysis 1 (AML disease state), and 4 sex-relevant DE genes were found in 3 different signaling pathways, including FLT3 and CD34 in Hematopoietic cell lineage, FLT3 in Transcriptional misregulation in cancer, and PMAIP1 in p53 signaling pathway, and down-regulated gene MS4A1 in Hematopoietic cell lineage.

Analysis 2b: Age-dependent gene expression meta-analysis and associated signaling pathways in AML compared to healthy individuals, from this analysis we found 372 DE unique age-dependent genes (375 DEPS) overlap with the 964 DE unique genes (974 DEPS) from our AML disease state meta-analysis (chapter 3 Fig. 19a), with 137 up- and 238 down- regulated. We also found 25 DE genes common across a baseline (0 to 19) age-group (chapter 3 Fig. 19g) with 15 genes were identified as potential therapeutic for drug target and 75 genes identified to be specific to one age-group (Fig. 19d) with 24 genes were categorized for “druggable genome”.

Finally, we used our results combined with a machine learning model (KNN model), and implemented supervised machine learning for classification training. We were able to test our model using 5 independent gene expression datasets (613 AML and healthy). Using

our trained model, we were able to classify AML patients compared to healthy individuals with > 90% achieved accuracy. Overall our findings provide a new reanalysis of public datasets, that enabled the identification of potential new gene sets relevant to AML that can potentially be used in future experiments and possible stratified disease diagnostics.

Outlook

While our results and analyses have identified important gene expression signatures relevant to AML, and many potential new drug-gene targets, our findings may generate more questions that should be considered in the future including, i) associations between age-groups and changes in gene expression across different AML subgroups to help improve AML risk stratification, ii) age-dependent pseudo time-series models to identify changes in gene expression with more specific AML patients age and sex. However, these questions and analysis would require many more well annotated AML patient's gene expression data that are currently unavailable, particularly given the heterogeneity of the disease. We hope new studies will address this in the future, and that our findings will lead to new findings that will help our understanding of AML, and ultimately improve disease diagnosis, prognosis and treatment.