Computational Biology-Driven Genomic and Epigenomic Delineation of Acute Myeloid Leukemia

Mathijs Sanders

ISBN: 978-94-6299-022-7

Cover design: Mathijs Sanders, with compliments to Blue Lightning TV and Ch-Ch-Check It Photoshop tutorials
Layout: Nikki Vermeulen, Ridderprint BV, Ridderkerk, The Netherlands
Printing: Ridderprint BV, Ridderkerk, The Netherlands

Alternative location supplemental material: http://www.planetmathematics.com/mathijs_sanders/

The work described in this thesis was performed at the Department of Hematology at the Erasmus University Medical Center, Rotterdam, The Netherlands. The work was funded by the Center for Translational Molecular Medicine.

Printing of this thesis was financially supported by: The Erasmus Unversity Rotterdam and the Center for Translational Molecular Medicine.

Computational Biology-Driven Genomic and Epigenomic Delineation of Acute Myeloid Leukemia

Computationeel biologisch gedreven genomische en epigenomische delineatie van acute myeloïde leukemie

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens het besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op
dinsdag 17 februari 2015 om 13:30 uur

door

Mathijs Arnoud Sanders

geboren te Rotterdam

PROMOTIECOMISSIE

Promotoren: Prof.dr. H.R. Delwel
Prof.dr. B. Löwenberg

Overige leden: Prof.dr. J.J. Goeman
Prof.dr. G.J. Ossenkoppele
Prof.dr. J.N.J. Philipsen

Copromotor: Dr. P.J.M. Valk

Voor mijn lieve Kristel en Jonathan
Voor mijn ouders

TABLE OF CONTENTS

CHAPTER

General introduction

Partially adapted from Curr Opin Hematol. 2013 Mar;20(2):79-85.

1. HEMATOPOIESIS AND LEUKEMIA

Hematopoiesis is the deterministic process of blood cell formation taking place in the bone marrow.[1] Mature blood cells are produced by a tightly controlled mechanism from hematopoietic stem cells (HSCs) residing in the bone marrow. Upon maturation blood cells are released into the peripheral blood and from this point onward can be transported to the different locations of the body. The mature blood cells exert different functions dependent on a strictly controlled path of maturation. The distinct leukocytes comprising granulocytes, monocytes, macrophages, natural killer cells and lymphocytes are essential for the defense against pathogens and foreign invaders, erythrocytes play a pivotal role in the transportation of oxygen to remote organs, and platelets confer the process of blood clotting.

Mature blood cells are short-lived and require continuous replenishment. The control of the production and the total number of blood cells is conferred by multipotent progenitors and a small population of pluripotent HSCs (Figure 1). HSCs reside in the bone marrow of adult mammals at the apex of a hierarchy of progenitors which become progressively restricted to several and eventually single lineages of blood cells.[2] Additionally these pluripotent stem cells have the unique ability to self-renew, generating a source for continuous replenishment of the complete blood cell system. The hematopoietic stem cell compartment contains stem cells with progressively decreased self-renewal capacity with the retention of multi-lineage reconstitution. The rare long term HSC (LT-HSC) is at the pinnacle of the hematopoietic hierarchy and is mainly quiescent. With the most conserved rate of self-renewal it prevents the depletion of the stem cell pool.[3] The less rare short term HSC (ST-HSC) still retains a minimal ability for self-renewal and is the more active effector cell for hematopoietic replenishment in normal situations.[4] The main constituent of the hematopoietic stem cell compartment is the multipotent progenitor (MPP) which lost its self-renewal capacity, however, kept the ability to give rise to daughter cells of different lineages. The daughter cells, common myeloid progenitor (CMP)[5] and common lymphoid progenitor (CLP)[6], are still oligopotent as they give rise to multiple blood cell types, e.g., lymphocytes, granulocytes, platelets and erythrocytes.

The production of mature blood cells is a strictly controlled process that adapts to the needs of human physiology, e.g., erythrocyte production after blood loss. The control is asserted mainly by external stimuli, e.g., hematopoietic cytokines or growth factors, which are produced by constituents of the regulatory microenvironment within the bone marrow niche, other blood cells or cytokine secreting organs.[7-9] The microenvironment plays a pivotal role in the formation of adequate numbers of blood cells of the correct type[10] and the hematopoietic cytokines it produces allows the hematopoietic system to dynamically adapt to extramedullary events, e.g., blood loss, infection or cancer immunoediting.[11]

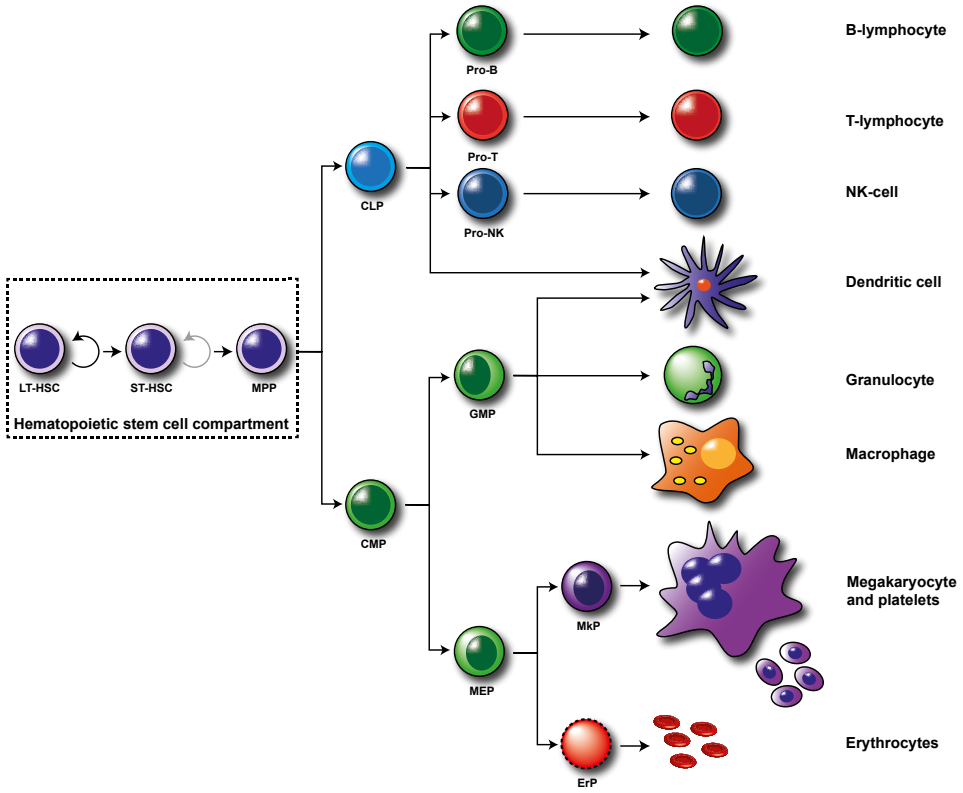


Figure 1. Hematopoiesis. (adapted from Reya et al.[3]) Within the hematopoietic stem cell compartment reside pluripotent stem cells with the ability to differentiate into any mature blood cell type. The hematopoietic stem cell compartment comprises three subtypes, at the apex the long term HSC (LT-HSC) which mainly self-renews, the short term HSC (ST-HSC) with self-renewal capacity and which transiently differentiates, the multi-potent progenitor (MPP) which has lost the self-renewal capacity but still gives rise to daughter cells of different lineages. The first single lineage progenitors are the common myeloid progenitor (CMP) and the common lymphoid progenitor (CLP). The CMP can give rise to the GMP; granulocyte myeloid precursor, MEP; megakaryocyte erythrocyte precursor, ErP; erythrocyte precursor, MkP; megakaryocyte precursor and finally the mature blood cells. The CLP gives rise to Pro-B, Pro-T and Pro-NK cells, which upon subsequent maturation give rise to B-cell, T-cell, and natural killer cell lymphocytes.

The malignant transformation of normal hematopoietic stem and progenitor cells (HSPC) results in the accumulation of hematopoietic cells in the bone marrow lacking the ability to differentiate with increased capacity for proliferation and survival. Sufficient accumulation of these non-functional hematopoietic cells impairs the function of the residual normal hematopoietic cells, eventually precluding the production of functional mature blood cells. The final outcome of this malignant process is termed leukemia and can be subdivided in chronic leukemia and acute leukemia. Chronic leukemia is characterized by an increased and unregulated production of white blood cells with differentiation capacity whereas acute leukemia is characterized by

the accumulation of the most immature hematopoietic stem and progenitor cells without differentiation capacity. Acute leukemia is a broad term for heterogeneous malignancies affecting the hematopoietic system and can be further subdivided based on the affected lineage, i.e., acute myeloid leukemia (AML)[12,13] and acute lymphoblastic leukemia (ALL).[14,15] The clinical distinction can be made on the basis of cell morphology, cell surface marker expression measured by immunohistochemical staining or specific gene expression patterns.[16,17]

2. ACUTE MYELOID LEUKEMIA

The focus of this thesis is primarily set on the genetic and epigenetic delineation of AML pathogenesis. AML is characterized by the accumulation of immature hematopoietic cells of the myeloid lineage in the bone marrow lacking the ability to differentiate towards functional granulocytes or monocytes and in rare cases also affecting the development of erythrocytes and megakaryocytes. The term AML specifies a broad spectrum of hematological malignancies and could be considered heterogeneous evidenced by the multitude of underlying abnormalities conferring variegated prognosis and response to therapy.

Epidemiology and clinical facets

AML is a rare disease with an incidence of approximately 3.8 cases per 100.000 individuals per year with a median age of 70 at presentation.[13] Additionally, myelodysplastic syndrome (MDS)[18] and myeloproliferative neoplasm (MPN)[19] are pre-leukemic disease entities which can progress towards AML. The first clinical symptoms observed at AML onset are infections, fatigue, hemorrhage and more rarely extramedullary involvement such as gingival hyperplasia, i.e., abnormal increased size of the gum, in cases of acute myelomonocytic or monoblastic leukemia.[20] The symptoms are the result of impaired normal hematopoiesis due to the accumulation of leukemic blasts in the bone marrow, precluding the production of functional mature blood cells. The dysfunction and aberrant distribution of malignant blood cells could readily explain the symptoms; infections (lack of granulocytes), fatigue (lack of erythrocytes), hemorrhage (lack of platelets), gingival hyperplasia (infiltration of malignant blood cells).

Treatment of AML is divided into an induction phase followed by a post-induction phase. The induction phase aims at eradicating the leukemic blasts by treatment with combinatorial intensive chemotherapy. Subsequent consolidation therapy is performed, when complete remission is received, aiming at the elimination of residual undetectable leukemic blasts by means of allogeneic or autologous stem cell transplantation or conventional chemotherapy. The type of consolidation therapy is highly dependent on a set of different clinical parameters, e.g., age, genetic markers and suitable stem cell donor.

The genetic and epigenetic landscape of AML

AML is a heterogeneous disease evidenced by the increasing number of cytogenetic and molecular abnormalities unveiled over the past decades. Traditionally, the classification of AML has been based on morphology, immunohistochemistry and immunophenotyping following the French-American-British (FAB) classification system.[21] The karyotype of leukemic blasts has been a pivotal prognostic marker for many years, although recent years have demonstrated that molecular genetic analyses are equally important, particularly for AML patients lacking any cytogenetic aberrations. In 2008 the AML classification model has been updated by the World Health Organization (WHO), which incorporates besides morphology also cytogenetics and molecular abnormalities.[22]

Cytogenetics in AML

A subgroup within this classification system comprises AML with recurrent cytogenetic abnormalities each with distinct clinical behavior and outcome (Table 1). Predominant among those are AML entities harboring cytogenetic abnormalities involving inv(16)(p13q22)/t(16;16)(p13;q22), t(8;21)(q22;q22), t(15;17)(q22;q12) or t(9;11)(p22;q21). These cytogenetic abnormalities are well characterized and could be further subdivided on the basis of additionally acquired abnormalities or gene expression markers.[23-26] Recent addendums to the classification model has introduced the AML entity with the cytogenetic abnormality inv(3)(q21q26.2) or t(3;3)(q21;q26.2), which results in the overexpression of the gene *ecotropic viral integration site-1* (*EVI1)* localized at 3q26.2. *Evi1* is identified as a common retroviral integration site in murine myeloid disorders.[27]

Table 1. Recurrent cytogenetic abnormalities in AML.

Cytogenetic abnormality	Frequency (%)	Genes involved	Prognostic significance
Normal karyotype	45	-	Intermediate
Complex karyotype	11	*TP53*	Unfavorable
+8	9	Unknown	Intermediate
t(15;17)(q22;q12)	8	*PML-RARA*	Favorable
-7/-7q	8	*CUX1*, *MLL3*	Unfavorable
-5/-5q	7	Unknown	Unfavorable
t(8;21)(q22;q22)	6	*RUNX1-ETO*	Favorable
inv(16)(p13q22)/t(16;16)(p13;q22)	5	*CBFB-MYH11*	Favorable
t/inv(11q23)	4	*KMT2A*	Favorable(*BRE*+) Unfavorable(*EVI1*+)
+21	3	Unknown	Intermediate
inv(3)(q21q26)/t(3;3)(q21;q26)	2	*EVI1*, *GATA2*	Unfavorable

Adapted from [29-32]

The protein EVI1 is a transcriptional regulator which invokes DNA interaction through its two zinc finger domains.[28] The cytogenetic abnormality does not result in a transcript fusion and the overexpression has long been postulated to be conferred by the repositioning of an enhancer element belonging to the gene *RPN1* located at 3q21. Hence the definition *RPN1-EVI1* for inv(3)(q21q26.2) or t(3;3)(q21;q26.2) malignancies in the WHO 2008 AML classification. In this thesis, and shown by others[29], we demonstrate that *EVI1* overexpression is conferred by the repositioning of a distal *GATA2* enhancer towards the human chromosome 3q26.2 region.

Mutational landscape and patterns in AML

The initiation of AML is not conferred by a single aberration as observed in the core binding factor (CBF) leukemias[24,33,34], i.e., AML with inv(16) or t(8;21) chromosomal abnormalities. Detailed studies have demonstrated that combinations of genetic alterations are necessary for the development of overt leukemia.[35,36] These alterations mainly perturb cellular mechanisms associated with differentiation, survival, apoptosis, self-renewal and proliferation. These genetic lesions can be dichotomized on size, e.g., large cytogenetic events (translocation, inversions, duplications, deletions, and amplifications), and small genetic lesions (mutations and small insertions and deletions). Recent efforts, in conjunction with a new technology called next generation sequencing (NGS), has brought to light the multitude of small recurrent genetic lesions acquired during leukemogenesis. These combinatorial mutational patterns reflect the heterogeneous nature of AML (Figure 2). Several of these recurrently acquired molecular abnormalities, such as mutations in the genes *nucleophosmin 1* (*NPM1*), *CCAAT enhancer binding protein alpha* (*CEBPA*), *fms-related tyrosine kinase 3* (*FLT3,* in particular internal tandem duplication [*FLT3*-ITD]), have independent prognostic values, especially in AML with normal cytogenetics (Table 2). The application of NGS has brought to light the existence of mutational patterns in AML.[35] Mutations in *additional sex combs-like 1* (*ASXL1*) were initially identified in MDS[37] and subsequently observed in AML. Mutations in *ASXL1* confer a dismal prognosis[38] and are inversely correlated with *NPM1* mutations and *FLT3*-ITD.[39] Recent efforts have led to the discovery of *GATA2* mutations in AML patients, more frequently in patients harboring biallelic mutations in *CEBPA*.[40] Strikingly, both mutations have been linked to familial predisposition for MDS or AML.[41,42]

Mutational mutual exclusivity: an example

Mutations in the gene *tet methyl-cytosine dioxygenase 2* (*TET2*)[43,44] were initially identified in MDS and subsequently detected in AML. The introduction of NGS led to the discovery of *isocitrate dehydrogenease* (*IDH*) mutations, i.e., *IDH1* and *IDH2*[45], conferring dismal prognosis in particular AML subtypes.[46] The protein TET2 plays an important role in the reversion of 5-methylcytosine (5-mC) towards ordinary cytosine and requires the cofactor alpha-ketoglutarate (α-KG) to exert its function. Mutations in this gene impairs the iterative hydroxylation of 5-mC resulting in its accumulation.[47] The cofactor α-KG is produced by IDH1 and IDH2 and mutations within each of the

genes encoding these proteins give rise to a neomorphic function, which confers α-KG processing into 2-hydroxyglutarate (2-HG). The oncometabolite 2-HG binds to TET2 and subsequently impairs its hydroxylation function.[48] Mutations in *TET2*, *IDH1* and *IDH2* are mutually exclusive and all perturb the hydroxylation of 5-mC by impairing TET2[48], providing an example of a mutually exclusive mutation pattern affecting the same pathway. Additionally, recurrent mutations in the gene *DNA (cytosine-5)-methyltransferase 3 alpha* (*DNMT3A*) were observed in AML.[49] The DNMT3A protein confers the *de novo* methylation of cytosines, implying that mutations in *DNMT3A*, *TET2*, *IDH1* and *IDH2* play a role in leukemogenesis by perturbing DNA methylation dynamics.

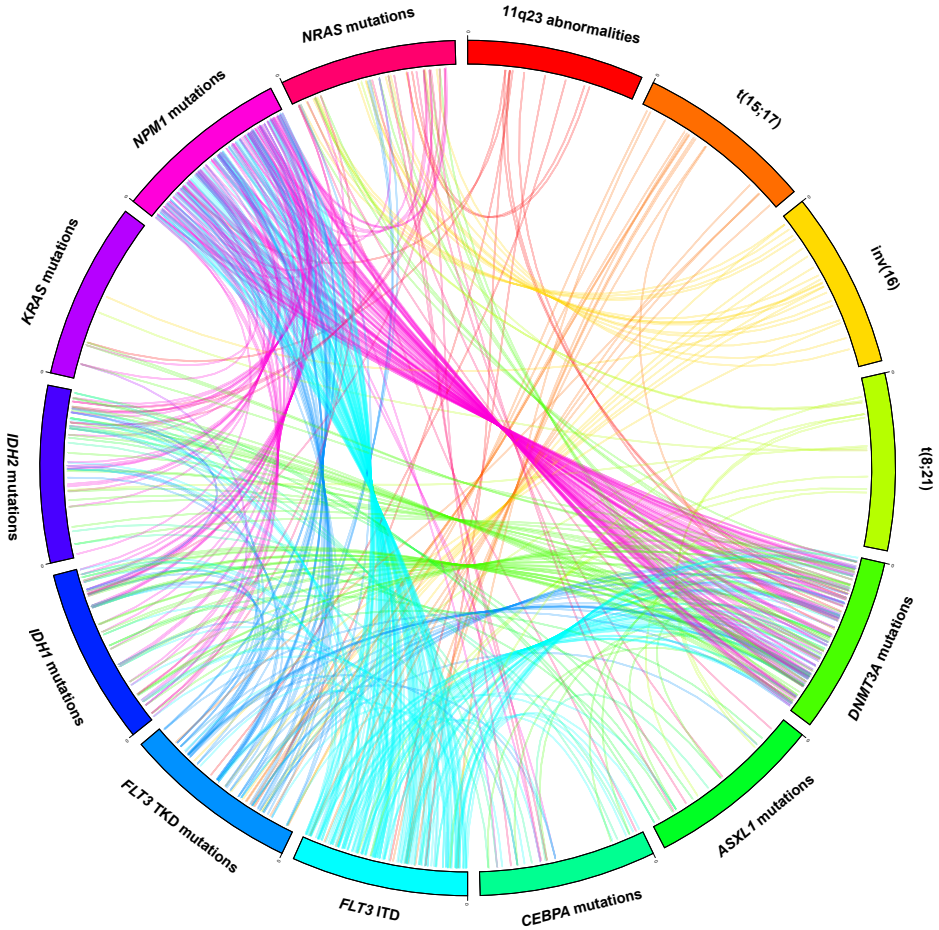


Figure 2. Molecular heterogeneity of AML. Circos plot illustrating the interrelationships of translocations and mutations found in a cohort of 498 *de novo* AML cases. Colored lines illustrate concurrent lesions found in one AML patient. Of note, patients can have more than two lesions.

Epigenetics

The field of epigenetics concerns the discipline investigating the dynamics of DNA conformation or accessibility. The DNA sequence serves as a blueprint for the production of functional messenger ribonucleic acid (mRNA), which in turn is translated into a protein exerting its functional properties according the necessities of the cell. Epigenetic alterations, e.g., DNA methylation[50], histone modification[51], chromatin looping[52,53], confers transcriptional control without changing the DNA sequence and could be considered as an additional layer of regulation by which the cell controls its requirements. Aberrant epigenetic patterns have been observed in AML[54,55] and adds a layer of complexity to the unraveling of AML pathogenesis. Recent NGS efforts have demonstrated that many epigenetic modifiers are recurrently mutated in AML[56], potentially resulting in aberrant epigenetic alterations.

Table 2. Recurrent molecular abnormalities in AML.

Gene symbol	Frequency (%)	Prognostic significance	Association with cytogenetics
ASXL1	5-10	Unfavorable	Normal
BCOR1	4	Undetermined	Normal
BCORL1	6	Undetermined	Normal
CBL	5	-	CBF AMLs
CEBPA	5-10	Favorable	Normal
Cohesin complex	5-10	-	Normal
DNMT3A	20-25	Unfavorable	Normal
FLT3-ITD	20	Unfavorable	Normal, t(15;17)
FLT3-TKD	5-10	Controversial	Normal
GATA2	2	Controversial	*CEBPA*, inv(3)/t(3;3)
IDH1	8	Unfavorable	Normal
IDH2	8	Unfavorable	Normal
KIT	2-8	Unfavorable	CBF AMLs
KRAS	5	-	inv(3)/t(3;3)
MLL-PTD	5-11 (CN-AML)	Unfavorable	Normal/Trisomy 11
NPM1	25-35	Favorable	Normal
NRAS	10-15	-	inv(16)/inv(3)/t(3;3)
RUNX1	10	Unfavorable	Normal/Trisomy 21
SF3B1	2-5	Undetermined	AML with ringed sideroblasts/ inv(3)/t(3;3)
TET2	20	Unfavorable	Normal
TP53	<10	Unfavorable	Complex karyotype (69%)
U2AF1	4	Undetermined	Secondary AML
WT1	10	Controversial	Normal

Adapted from [57-61]

3. ACUTE LYMPHOBLASTIC LEUKEMIA

A single chapter of this thesis is dedicated to ALL and therefore mandates a brief introduction into this hematological malignancy. ALL is characterized by the accumulation of primitive lymphoblastic cells in the bone marrow that have lost the ability to differentiate towards functional T-cells, B-cells or NK-cells. Within each respective lineage (Figure 1) there is still a heterogeneous group of disorders with variegated underlying genetic abnormalities and clinical behavior. This thesis only focusses on adult B-cell ALL (B-ALL) and T-cell ALL (T-ALL).

Epidemiology

The incidence of ALL is approximately 1 case per 50.000 individuals.[14,15] Pediatric leukemia accounts for approximately 70% of ALL cases and is the most common childhood cancer. The incidence of ALL peaks at 0-5 years, while in the adult cases the incidence peaks in patients older than 65 years. Clinical symptoms range from general weakness and fatigue to anemia, bone pain, and enlarged lymph nodes and spleen. The five-year event-free survival is nearly 80 percent for children and approximately 40 percent for adults.[62,63]

Cytogenetic and genetic lesions

Cytogenetics and the mutational landscape of ALL

Cytogenetics and cytology play a pivotal role in the classification of ALL. Although a FAB classification exists for ALL, it is far less used than for AML. First, ALL is stratified according to cytomorphology or cell surface markers into B-ALL or T-ALL. Secondly, cytogenetics is used for the determination of recurrent chromosomal aberrations. Most of these cytogenetic abnormalities are the result of translocations leading to the expression of oncogenic fusion transcripts (Table 3). Specific B-ALL aberrations comprise: t(12;21)(p13;q22) (*ETV6-RUNX1*), t(1;19)(q23;p13.3) (*TCF3-PBX1*), t(9;22)(q34;q11.2) (*BCR-ABL1*) and *MLL*-rearrangements. Specific T-ALL aberrations comprise: del(1p32) (*SIL-TAL1*). Prognostication in adult ALL is limited to a few prognostic genetic markers, e.g., *BCR-ABL1* and *MLL-AF4*. Traditional non-genetic markers, such as age, sex and WBC, have a significant effect on treatment outcome. Numerical changes of chromosomes has been determined to affect treatment outcome: strong hyperdiploidy (more than 50 chromosomes) results in a favorable prognosis status and normal or low hypoploidy results in an intermediate and unfavorable prognosis status, respectively.[64] The presence of the BCR-ABL fusion results in a dismal prognosis and is commonly observed in adult ALL. The determined ALL mutational landscape revealed specific mutations in different ALL subtypes: (I) frequent *DNMT3A* mutations in adult early thymocytes progenitor-ALL (ETP-ALL)[65], (II) frequent *tumor protein P53* (*TP53*), *IKAROS family zinc finger 2* (*IKZF2*) and *retinoblastoma 1* (*RB1)* lesions in pediatric hypodiploid ALL[66], (III) kinase-activating lesions in *BCR-ABL1*-like ALL.[67]

Table 3. Recurrent cytogenetic abnormalities in ALL.

Cytogenetic abnormality	Frequency (%)	Genes involved	Prognostic significance
Normal karyotype	15-34	-	Intermediate
t(9;22)(q34;q11.2)	11-29	*BCR-ABL1*	Unfavorable
t(4;11)(q21;q23)	4-9	*KMT2A-AFF1*	Unfavorable
Hyperdiploidy	7-8	-	Favorable/intermediate
Hypodiploidy	7-8	-	Unfavorable
t(1;19)(q23;p13.3)	1-6	*TCF3-PBX1*	Intermediate
t(12;21)(p13;q22)	0-4	*ETV6-RUNX1*	Undetermined
SIL-SCL deletion	3	*SIL-TAL1*	Unfavorable

Adapted from [64]

Recurrent focal genetic lesions in ALL

Beside recurrent cytogenetic abnormalities and small mutations, ALL is characterized by recurrent deletions and amplifications perturbing single or multiple genes in both pediatric and adult ALL.[68,69] A small number of novel and recurrent genetic lesions have been demonstrated to be acquired in B-ALL and T-ALL, e.g., the deletion of *cyclin dependent kinase inhibitor 2A* and *2B* (*CDKN2A/B)* and the deletion of *RB1*. The remainder of recurrent genetic lesions are specific for B-ALL; deletion of *IKAROS family zinc finger 1* (*IKZF1*), *paired box 5* (*PAX5*), *B and T lymphocyte associated* (*BTLA*), while others are more specific for T-ALL; deletion of *Wilms tumor 1* (*WT1*) and *neurofibromin 1* (*NF1*). These genetic lesions are observed in varying degrees within pediatric and adult ALL, e.g., *IKZF1* deletions are more prominent in adult than pediatric ALL cases lacking the t(9;22) cytogenetic aberration.[70] The underlying mechanism for these deletions has long been debated to be associated with illegitimate RAG-mediated rearrangements.[71] The adaptive immune system requires diversification in defense of pathogens or other foreign invaders. It maintains this diversity by generating a vast repertoire of antigens by combining the antigen constituents in a combinatorial manner.[72] The antigen constituents comprise variable (V), diversity (D) and joining (J) gene segments in the antigen receptor gene regions. This recombination process is mediated by the recombination activating gene (RAG) proteins, i.e., RAG1 and RAG2, which recognize recombination signal sequences (RSS) flanking the gene segments. The existence of cryptic RSS flanking deletion breakpoints has led to the hypothesis that deletion events are invoked by the RAG complex[71,73] and recently been shown to be a prominent driver of rearrangements in *ETV6-RUNX1* ALL cases.[74]

4. GENOME-WIDE APPROACHES FOR THE DELINEATION OF AML

Recent technological advances have allowed for the genome-wide characterization of AML. This progression can be explained by the improvement and flexibility of novel experimental

tools, the quick procession of measuring devices and statistical or algorithmic developments. Detailing each genome-wide characterization technology is warranted for the discernment of its advantages and caveats.

Gene expression profiling

The human genome contains thousands of genes and their products, i.e., mRNA and proteins, function within a complicated web of biological mechanisms. The attainment of gene expression levels allows the researcher to unravel this complex web of interactions. In human disease, gene expression level assessment enables the determination of aberrant gene expression patterns potentially allowing for the further subcategorization of established AML entities or delineation of aberrant mechanisms affiliated to leukemogenesis. Technological advances in the past decades have led to a major breakthrough enabling the genome-wide characterization of tumor material by array-based technologies. First among these array-based technologies was the gene expression array that allows for the genome-wide measurement of the human transcriptome (Figure 3A). This genome-wide technique measures the level of thousands of mRNA transcripts simultaneously by a process called gene expression profiling (GEP). The glass slide of the array is spotted by thousands of DNA probes that can, based on their sequence specificity, competitively hybridize to complementary cDNA/cRNA produced from mRNA. The rate of hybridization is measured and used for the estimation of gene expression levels. These expression profiles have several biologically and clinical relevant applications. Initially, GEPs were utilized for the identification of different AML subtypes[17,75,76] and were pivotal in cementing the homogeneity of AML entities with recurrent cytogenetic abnormalities, i.e., t(15;17), t(8;21) and inv(16)/t(16;16), as well as the identification of novel AML subtypes with specific gene expression patterns.[77-79] Prognostic expression markers for clinical purposes can be discerned from GEPs for further classification of AML.[26,80-82] Importantly, GEP could give insight into the biological mechanisms perturbed by the underlying genetic abnormality.[78] A multitude of relevant applications can be devised for the delineation of AML pathogenesis by GEP, however, it is limited to the single facet of gene expression levels and is therefore unable to identify all aberrant processes.

DNA mapping arrays

The gain or loss of chromosomal regions may result in the perturbation of AML initiating genes. Traditionally, cytogenetics was employed to detect large chromosomal abnormalities, but was limited in its resolution and therefore unable to detect smaller genetic alterations. DNA mapping arrays, likewise to gene expression arrays, are utilized by means of hybridization procedures. The glass slide is spotted by DNA probes that can, based on their sequence specificity, bind to particular segments of DNA. Specific probes are generated to measure the genotype of single nucleotide polymorphisms (SNPs), which are variants in the genome observed in at least some percentage of the healthy human population. One probe measures the 'A' allele of a specific DNA segment

while another slightly different probe measures the 'B' allele for the same segment. The amount of hybridization can be used for the determination of the copy number when compared to an appropriate base line control. In a normal situation the copy number equals 2, one chromosome from the father and the other chromosome of the mother. The loss or gain of DNA material leads to copy number variations (CNVs) and have been implicated in oncogenesis (Figure 3C).[83] With an appropriate control, e.g., normal tissue or remission material, the loss of heterozygous SNP genotypes, i.e., loss-of-heterozygosity (LOH), can highlight regions with cancer-critical genes.[84] In general AML is characterized by the frequent acquisition of LOH in comparison to the low number of observed CNVs[85], which are more frequently observed in ALL.[68,69]

Next generation sequencing

Technical advances of the last decade have led to the introduction of NGS. Different NGS methodologies take as input RNA or DNA derivatives and determine the sequence of millions of DNA fragments simultaneously in a manner reminiscent to Sanger sequencing. These fragments can be partially sequenced from one side, called single-end sequencing, or partially sequenced from both sides, called paired-end sequencing. Different applications for NGS have hitherto been developed, e.g., whole genome sequencing (WGS), whole exome sequencing (WES), (m)RNA sequencing (RNA-Seq), chromatin conformation capture sequencing (4C-Seq), and chromatin immunoprecipitation sequencing (Chip-Seq). These techniques result in millions of reads and alignment of these reads to the genome of interest allows for the quantification of processes of interest, e.g., gene expression levels, or the identification of genetic lesions (Figure 3C). Due to its high sensitivity and accuracy NGS rapidly replaces array based technologies.[86] All mentioned NGS techniques, *vide supra*, will be employed during this thesis and therefore mandates a brief introduction detailing its use, benefits and pitfalls.

Whole exome and genome sequencing

The determination of the complete DNA sequence can be achieved by WGS. DNA from the tissue of interest is isolated and preprocessed without any form of sequence enrichment. Contemporary NGS technology allows for the sequencing of the complete genome, albeit with a low coverage. The coverage implies the number or depth of DNA or RNA fragments that have been sequenced for a particular region. A higher coverage increases detection power of somatic mutations or a better estimation of quantities of interest. Subsequently, WGS is not preferable when mutations could be expected to be present in only a fraction of leukemic blasts. A higher coverage is achieved by selectively isolating DNA regions of interest, e.g., the exome, and are procured by target enrichment or exome capture, before sequencing is performed. WES supersedes WGS at detecting mutations within the exome or determining the clonal architecture of AML. The caveat of capture procedures concerns the accidental capture of regions with high homology, sometimes resulting in biased estimations or false positive mutations.

Whole transcriptome sequencing

RNA-Seq quantifies the gene expression levels reminiscent to gene expression arrays (Figure 3A). RNA-Seq requires no *a priori* knowledge about the transcriptome which is needed for array-based technologies. The number of fragments aligning to a gene, corrected for gene length and the total number of sequenced fragments, is used for gene expression level estimation.[87] The mRNA transcripts originating from a gene can have different forms owing to gene isoforms or RNA splicing and could possibly encode proteins with vastly different functions.[88] Additionally, RNA-Seq can be used for the detection of fusion transcripts[89] or mutations in genes if they are expressed.[90] Finally, the accurate detection of gene expression levels is highly correlated with the coverage.

Chip-Seq

Chip-Seq is used for mainly two reasons. First, Chip-Seq can be used for the identification of protein-DNA interactions and to quantify the frequency of this interaction (Figure 3B). Second, it allows for the detection of epigenetic alterations, e.g., histone modifications. The protein of interest is cross-linked with DNA in viable cells and immunoprecipitated.[91] The isolated DNA fragments are sequenced and the resultant reads are aligned against the genome of interest. Subsequently an interaction profile is generated from the aligned reads representing the interaction frequency of the protein of interest with particular DNA segments. Gene expression is a tightly controlled process that is modulated by the binding of proteins to functional genomic elements affiliated with the gene. Classically, these genomic elements comprise promoters, enhancers, insulators or CpG islands. Recently, the term super-enhancers is introduced to signify highly active epigenetic regions comprising clusters of enhancers and have been associated with cell identity and disease.[92,93] Chip-Seq allows for the identification of these genomic elements and the proteins that bind to them (Figure 3B).

4C-Seq

4C-Seq quantifies the frequency of two DNA segments being in close proximity or interacting.[94] Chromatin loops, under the direction of the cohesion complex, bring specific genomic elements in proximity, e.g., the promoter of a gene and an affiliated enhancer, invoking transcriptional control.[95] In brief, at first a region of interest is chosen and polymerase chain reaction (PCR) primers are designed accordingly. DNA segments in close proximity are cross-linked in viable cells and DNA is fragmented with a restriction enzyme. Subsequently, the ends of the fragmented cross-linked DNA are ligated and cross-linking is removed resulting in circular DNA. Further processing with a second restriction enzyme is enacted and PCR amplification is performed with the designed primers. These primers are specific for the region of interest and therefore only amplifies the interacting DNA segment. These amplified fragments are subsequently sequenced and aligned. Interaction profiles are extracted and the density of interaction fragments in a region of interest relates to the interaction frequency.

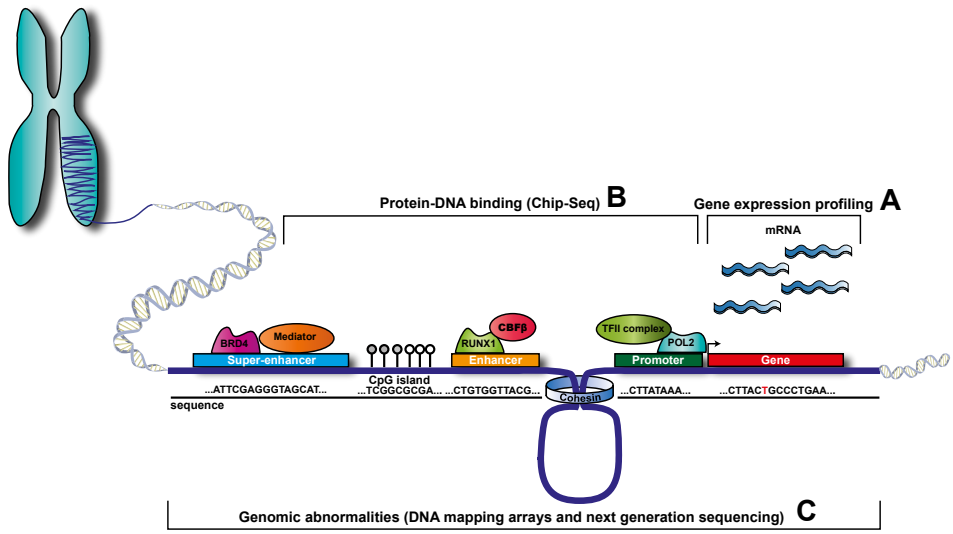


Figure 3. Genome-wide approaches in human disease. DNA is the blueprint of life and its sequence contains a large number of functional genomic elements. (A) The transcription of a gene results in the production of mRNA and its expression level can be determined by gene expression profiling. (B) The expression of genes is under control of many different genomic elements, e.g., promoters, enhancers, CpG islands and super-enhancers, on which many different proteins can bind to exert mRNA level control and this binding can be ascertained by Chip-Seq. (C) Perturbations of the genetic material is a hallmark of cancer and changes can be ascertained by DNA mapping arrays or next-generation sequencing. DNA helix was generated by the Illustrator DNA brush of James Hedberg.

5. SCOPE AND OUTLINE OF THE THESIS

The use of different genome-wide analysis approaches has provided avenues for further delineation of AML and ALL pathogenesis. This thesis focuses on the use and extension of these approaches to provide further insight into these diseases on a genetic and epigenetic level. The analytical tools for processing the results from these approaches have thus far been limited and therefore novel statistical models have been devised, focusing on their applicability to the procured data. Additionally, the functional and prognostic role of the novel detected genetic aberrations are further investigated.

The thesis consists out of three parts related to the different types of genome-wide approaches used. The first part (**chapter 2, 3**) concerns research utilizing gene expression profiling for prognostication or AML subtype classification. Classification methodologies for gene expression data are limited and unable to provide interpretable prediction signatures. This question is addressed in **chapter 2** and led to the development of the group lasso penalization framework for the multinomial logistic regression model. In **chapter 3**, we determine if AML patients harboring biallelic mutations in the gene *CEBPA* (*CEBPA*^{dm}) have distinct prognostic outcomes and GEPs. The

multinomial logistic regression model with lasso penalization is used for determining the gene expression signature of CEBPA^{dm} AML patients. We demonstrate in a validation cohort that our gene expression signature can perfectly discern CEBPA^{dm} cases from all other AML cases.

The second part (**chapter 4, 5**) focuses on the use of DNA mapping arrays to determine recurrent genetic alterations in AML and ALL. In **chapter 4**, we describe a software package for visualizing CNV profiles in conjunction with GEPs. In **chapter 5**, we identify recurrent genetic lesions in ALL and AML using DNA mapping arrays. We report on the recurrence of particular genetic lesions and with the use of NGS we demonstrate that many of the genetic alterations are the result of illegitimate RAG-mediated rearrangements in particular ALL subtypes.

The third part (**chapter 6, 7, 8, 9, 10, 11**) focuses on the use of NGS to address a multitude of research questions concerning the delineation of AML pathogenesis. The first five chapters focus on the determination of mutations or genetic aberrations in AML, while the last chapter describes a statistical framework for the determination of CNVs from NGS data. In **chapter 6**, we performed a mutational time-series analysis on pre-leukemic or leukemic material from a patient with severe congenital neutropenia who progressed towards AML with substantial delay. In **chapter 7**, we focus on understanding the underlying leukemogenic mechanism of inv(3)(q21q26) and t(3;3)(q21;q26) AML entities. Initially, we determined all the breakpoints in the 3q21 and 3q26 loci and discerned an asymptotic pattern of breakpoints in 3q21. The integration of RNA-Seq, Chip-Seq and 4C-Seq revealed that *EVI1* overexpression is the result of the repositioning of a distal *GATA2* enhancer towards the 3q26 locus, concurrently resulting in the hemizygous and reduced expression of *GATA2*. In **chapter 8**, we determine mutational patterns in the inv(3)(q21q26) and t(3;3)(q21;q26) AML subtypes. We reveal the predominant presence of activating RAS/RTK mutations in 98% of the cases and additionally reveal recurrent mutations in *GATA2*, *SF3B1* and *RUNX1*. In **chapter 9,** we utilized RNA-Seq to discover a previously unreported *KMT2A-MYH11* fusion transcript. In **chapter 10,** we used targeted resequencing to identify jumping translocations involving the gene *BCL11B*. We demonstrate that the jumping translocation integrates into super-enhancers with subsequent overexpression of *BCL11B*. In **chapter 11**, we develop a new algorithm that can determine CNV profiles from WGS and WES data. Algorithms determining CNVs from NGS data are still lacking or do not employ valuable noise statistics derived from a NGS reference set of diploid cases. We demonstrate that we can attenuate systematic bias and artifacts conferred by repeat regions on a set of AML cases characterized by WGS or WES. Finally, the results presented in this thesis are summarized and discussed in **chapter 12**.

REFERENCES

1. Orkin SH, Zon LI. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*. 2008;132(4):631-644.
2. Orkin SH. Diversification of haematopoietic stem cells to specific lineages. *Nat Rev Genet*. 2000;1(1):57-64.
3. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature*. 2001;414(6859):105-111.
4. Sharpless NE, DePinho RA. How stem cells age and why this makes us grow old. *Nat Rev Mol Cell Biol*. 2007;8(9):703-713.
5. Akashi K, Traver D, Miyamoto T, Weissman IL. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*. 2000;404(6774):193-197.
6. Kondo M, Weissman IL, Akashi K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell*. 1997;91(5):661-672.
7. Zhang J, Niu C, Ye L, et al. Identification of the haematopoietic stem cell niche and control of the niche size. *Nature*. 2003;425(6960):836-841.
8. Shen Y, Nilsson SK. Bone, microenvironment and hematopoiesis. *Curr Opin Hematol*. 2012;19(4):250-255.
9. Lotem J, Sachs L. Cytokine control of developmental programs in normal hematopoiesis and leukemia. *Oncogene*. 2002;21(21):3284-3294.
10. Raaijmakers MH, Mukherjee S, Guo S, et al. Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. *Nature*. 2010;464(7290):852-857.
11. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol*. 2002;3(11):991-998.
12. Lowenberg B, Downing JR, Burnett A. Acute myeloid leukemia. *N Engl J Med*. 1999;341(14):1051-1062.
13. Estey E, Dohner H. Acute myeloid leukaemia. *Lancet*. 2006;368(9550):1894-1907.
14. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. *N Engl J Med*. 2004;350(15):1535-1548.
15. Pui CH, Robison LL, Look AT. Acute lymphoblastic leukaemia. *Lancet*. 2008;371(9617):1030-1043.
16. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*. 1999;286(5439):531-537.
17. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1617-1628.
18. Hofmann WK, Koeffler HP. Myelodysplastic syndrome. *Annu Rev Med*. 2005;56:1-16.
19. Tefferi A, Vainchenker W. Myeloproliferative neoplasms: molecular pathophysiology, essential clinical understanding, and treatment strategies. *J Clin Oncol*. 2011;29(5):573-582.
20. Demirer S, Ozdemir H, Sencan M, Marakoglu I. Gingival hyperplasia as an early diagnostic oral manifestation in acute monocytic leukemia: a case report. *Eur J Dent*. 2007;1(2):111-114.
21. Bennett JM, Catovsky D, Daniel MT, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br J Haematol*. 1976;33(4):451-458.
22. Campo E, Swerdlow SH, Harris NL, Pileri S, Stein H, Jaffe ES. The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood*. 2011;117(19):5019-5032.
23. Schnittger S, Bacher U, Haferlach C, Kern W, Alpermann T, Haferlach T. Clinical impact of FLT3 mutation load in acute promyelocytic leukemia with t(15;17)/PML-RARA. *Haematologica*. 2011;96(12):1799-1807.
24. Care RS, Valk PJ, Goodeve AC, et al. Incidence and prognosis of c-KIT and FLT3 mutations in core binding factor (CBF) acute myeloid leukaemias. *Br J Haematol*. 2003;121(5):775-777.
25. Bindels EM, Havermans M, Lugthart S, et al. EVI1 is critical for the pathogenesis of a subset of MLL-AF9-rearranged AMLs. *Blood*. 2012;119(24):5838-5849.
26. Groschel S, Schlenk RF, Engelmann J, et al. Deregulated expression of EVI1 defines a poor prognostic subset of MLL-rearranged acute myeloid leukemias: a study of the German-Austrian Acute Myeloid Leukemia Study Group and the Dutch-Belgian-Swiss HOVON/SAKK Cooperative Group. *J Clin Oncol*. 2013;31(1):95-103.

27. Mucenski ML, Taylor BA, Ihle JN, et al. Identification of a common ecotropic viral integration site, Evi-1, in the DNA of AKXD murine myeloid tumors. *Mol Cell Biol*. 1988;8(1):301-308.

28. Nucifora G, Laricchia-Robbio L, Senyuk V. EVI1 and hematopoietic disorders: history and perspectives. *Gene*. 2006;368:1-11.

29. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415-427.

30. Mrozek K, Heerema NA, Bloomfield CD. Cytogenetics in acute leukemia. *Blood Rev*. 2004;18(2):115-136.

31. McNerney ME, Brown CD, Wang X, et al. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. *Blood*. 2013;121(6):975-983.

32. Chen C, Liu Y, Rappaport AR, et al. MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. *Cancer Cell*. 2014;25(5):652-665.

33. Valk PJ, Bowen DT, Frew ME, Goodeve AC, Lowenberg B, Reilly JT. Second hit mutations in the RTK/RAS signaling pathway in acute myeloid leukemia with inv(16). *Haematologica*. 2004;89(1):106.

34. Boissel N, Leroy H, Brethon B, et al. Incidence and prognostic impact of c-Kit, FLT3, and Ras gene mutations in core binding factor acute myeloid leukemia (CBF-AML). *Leukemia*. 2006;20(6):965-970.

35. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.

36. Frohling S, Scholl C, Gilliland DG, Levine RL. Genetics of myeloid malignancies: pathogenetic and clinical implications. *J Clin Oncol*. 2005;23(26):6285-6295.

37. Gelsi-Boyer V, Trouplin V, Adelaide J, et al. Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br J Haematol*. 2009;145(6):788-800.

38. Pratcorona M, Abbas S, Sanders MA, et al. Acquired mutations in ASXL1 in acute myeloid leukemia: prevalence and prognostic value. *Haematologica*. 2012;97(3):388-392.

39. Patel JP, Gonen M, Figueroa ME, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med*. 2012;366(12):1079-1089.

40. Greif PA, Dufour A, Konstandin NP, et al. GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. *Blood*. 2012;120(2):395-403.

41. Pabst T, Eyholzer M, Haefliger S, Schardt J, Mueller BU. Somatic CEBPA mutations are a frequent second event in families with germline CEBPA mutations and familial acute myeloid leukemia. *J Clin Oncol*. 2008;26(31):5088-5093.

42. Hahn CN, Chong CE, Carmichael CL, et al. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet*. 2011;43(10):1012-1017.

43. Delhommeau F, Dupont S, Della Valle V, et al. Mutation in TET2 in myeloid cancers. *N Engl J Med*. 2009;360(22):2289-2301.

44. Langemeijer SM, Kuiper RP, Berends M, et al. Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat Genet*. 2009;41(7):838-842.

45. Ward PS, Patel J, Wise DR, et al. The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell*. 2010;17(3):225-234.

46. Abbas S, Lugthart S, Kavelaars FG, et al. Acquired mutations in the genes encoding IDH1 and IDH2 both are recurrent aberrations in acute myeloid leukemia: prevalence and prognostic value. *Blood*. 2010;116(12):2122-2126.

47. Ko M, Huang Y, Jankowska AM, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature*. 2010;468(7325):839-843.

48. Figueroa ME, Abdel-Wahab O, Lu C, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell*. 2010;18(6):553-567.

49. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med*. 2010;363(25):2424-2433.

50. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*. 2010;11(3):204-220.

51. Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet*. 2009;10(5):295-304.

52. Kadauke S, Blobel GA. Chromatin loops in gene regulation. *Biochim Biophys Acta*. 2009;1789(1):17-25.

53. Sexton T, Bantignies F, Cavalli G. Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. *Semin Cell Dev Biol*. 2009;20(7):849-855.

54. Figueroa ME, Lugthart S, Li Y, et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*. 2010;17(1):13-27.

55. Taskesen E, Havermans M, van Lom K, et al. Two splice-factor mutant leukemia subgroups uncovered at the boundaries of MDS and AML using combined gene expression and DNA-methylation profiling. *Blood*. 2014;123(21):3327-3335.

56. Abdel-Wahab O, Levine RL. Mutations in epigenetic modifiers in the pathogenesis and therapy of acute myeloid leukemia. *Blood*. 2013;121(18):3563-3572.

57. Marcucci G, Haferlach T, Dohner H. Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *J Clin Oncol*. 2011;29(5):475-486.

58. Grossmann V, Tiacci E, Holmes AB, et al. Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood*. 2011;118(23):6153-6163.

59. Li M, Collins R, Jiao Y, et al. Somatic mutations in the transcriptional corepressor gene BCORL1 in adult acute myelogenous leukemia. *Blood*. 2011;118(22):5914-5917.

60. Je EM, Yoo NJ, Kim YJ, Kim MS, Lee SH. Mutational analysis of splicing machinery genes SF3B1, U2AF1 and SRSF2 in myelodysplasia and other common tumors. *Int J Cancer*. 2013;133(1):260-265.

61. Kon A, Shih LY, Minamino M, et al. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat Genet*. 2013;45(10):1232-1237.

62. Schrappe M, Reiter A, Ludwig WD, et al. Improved outcome in childhood acute lymphoblastic leukemia despite reduced use of anthracyclines and cranial radiotherapy: results of trial ALL-BFM 90. German-Austrian-Swiss ALL-BFM Study Group. *Blood*. 2000;95(11):3310-3322.

63. Linker C, Damon L, Ries C, Navarro W. Intensified and shortened cyclical chemotherapy for adult acute lymphoblastic leukemia. *J Clin Oncol*. 2002;20(10):2464-2471.

64. Mrozek K, Harper DP, Aplan PD. Cytogenetics and molecular genetics of acute lymphoblastic leukemia. *Hematol Oncol Clin North Am*. 2009;23(5):991-1010, v.

65. Neumann M, Heesch S, Schlee C, et al. Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. *Blood*. 2013;121(23):4749-4752.

66. Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet*. 2013;45(3):242-252.

67. Roberts KG, Li Y, Payne-Turner D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N Engl J Med*. 2014;371(11):1005-1015.

68. Mullighan CG, Goorha S, Radtke I, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007;446(7137):758-764.

69. Okamoto R, Ogawa S, Nowak D, et al. Genomic profiling of adult acute lymphoblastic leukemia by single nucleotide polymorphism oligonucleotide microarray and comparison to pediatric acute lymphoblastic leukemia. *Haematologica*. 2010;95(9):1481-1488.

70. Tokunaga K, Yamaguchi S, Iwanaga E, et al. High frequency of IKZF1 genetic alterations in adult patients with B-cell acute lymphoblastic leukemia. *Eur J Haematol*. 2013;91(3):201-208.

71. Waanders E, Scheijen B, van der Meer LT, et al. The origin and nature of tightly clustered BTG1 deletions in precursor B-cell acute lymphoblastic leukemia support a model of multiclonal evolution. *PLoS Genet*. 2012;8(2):e1002533.

72. Gellert M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu Rev Biochem*. 2002;71:101-132.

73. Mullighan CG, Phillips LA, Su X, et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*. 2008;322(5906):1377-1380.

74. Papaemmanuil E, Rapado I, Li Y, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet*. 2014;46(2):116-125.

75. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*. 2009;94(1):131-134.

76. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1605-1616.

77. Wouters BJ, Lowenberg B, Erpelinck-Verschueren CA, van Putten WL, Valk PJ, Delwel R. 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*. 2009;113(13):3088-3091.

78. Verhaak RG, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood*. 2005;106(12):3747-3754.

79. Hollink IH, van den Heuvel-Eibrink MM, Arentsen-Peters ST, et al. NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. *Blood*. 2011;118(13):3645-3656.

80. Eppert K, Takenaka K, Lechman ER, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med*. 2011;17(9):1086-1093.

81. Metzeler KH, Hummel M, Bloomfield CD, et al. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood*. 2008;112(10):4193-4201.

82. Rockova V, Abbas S, Wouters BJ, et al. Risk stratification of intermediate-risk acute myeloid leukemia: integrative analysis of a multitude of gene mutation and gene expression markers. *Blood*. 2011;118(4):1069-1076.

83. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med*. 2009;1(6):62.

84. Thiagalingam S, Laken S, Willson JK, et al. Mechanisms underlying losses of heterozygosity in human colorectal cancers. *Proc Natl Acad Sci U S A*. 2001;98(5):2698-2702.

85. Radtke I, Mullighan CG, Ishii M, et al. Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proc Natl Acad Sci U S A*. 2009;106(31):12944-12949.

86. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet*. 2008;24(3):133-141.

87. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5(7):621-628.

88. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*. 2013;497(7447):127-131.

89. McPherson A, Hormozdiari F, Zayed A, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol*. 2011;7(5):e1001138.

90. Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res*. 2009;37(16):e106.

91. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol*. 2008;26(12):1351-1359.

92. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell*. 2013;155(4):934-947.

93. Loven J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*. 2013;153(2):320-334.

94. van de Werken HJ, Landan G, Holwerda SJ, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods*. 2012;9(10):969-972.

95. Kagey MH, Newman JJ, Bilodeau S, et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*. 2010;467(7314):430-435.

CHAPTER 2

Sparse multi-class prediction based on the group lasso in multinomial logistic regression

Mathijs A. Sanders[1,2] and Jelle J. Goeman[2,3]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands

[2] Leiden University Medical Center, Department of Medical Statistics and Bioinformatics, Leiden, The Netherlands

[3] Radboud University Medical Center, Department for Health Evidence, Nijmegen, The Netherlands

Work in progress

ABSTRACT

Continuous variable selection using shrinkage procedures have recently been considered as favorable models in a wide range of scientific research; in particular biomedical research. In some cases, it is desirable to select as few predictors as possible, to increase the interpretability of the attained prediction rule. One frequently used shrinkage procedure; the lasso, imposes a L_1 regularization on the regression coefficients of general linear models, inherently leading to sparse prediction rules. For multi-class prediction in generalized linear models each predictor has a regression coefficient for each class. A major disadvantage is that the lasso selects individual regression coefficients instead of the more logical selection of complete predictors. Here we demonstrate a new regularization procedure, based on the group lasso in the multinomial logistic regression model. This methodology results in a lower number of retained predictors, but with similar prediction accuracy when compared to the lasso regularization scheme. To illustrate the new regularization applicability we have employed it on a large cohort of acute myeloid leukemia patients (AML, n=540) who are characterized on a gene expression microarray.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter2/

INTRODUCTION

Regression models put an emphasis on determining explanatory variables which predict response variables accurately. Contemporary high-throughput technologies, have given rise to vast amounts of high-dimensional data. Given the high-dimensionality of the data, it is worthwhile to perform variable selection, as sparser prediction rules are generally more interpretable and give more prediction power. Best-subset procedures are in most cases computationally intensive; even for a moderate number of variables, and are known to be unstable due to their discrete nature.[1] More robust strategies have been proposed for the multinomial logistic regression model[2-4] by imposing a penalty on the regression coefficients.[5] In these logistic regression models each class has its own set of regression coefficients and imposing penalizations on these coefficients confers modeling constraints. The lasso penalization scheme[6] puts a L_1 regularization on the regression coefficients. If predictors are pair-wise correlated, e.g., genes co-regulated, the penalization scheme will only retain one predictor, discarding the remainder of correlated predictors. In the usual logistic regression setting we have a continuous response $\boldsymbol{Y} \in \mathbb{R}^n$, a $n \times p$ design matrix X, a regularization parameter λ and a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^p$. The lasso estimator is defined as:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (1)$$

For large values of λ, some coefficients of the estimator $\hat{\boldsymbol{\beta}}$ become exactly zero and are considered unassociated to the response variable. The lasso procedure provides sparse prediction rules, a beneficial feature when utilizing high-dimensional data.[7] In a multi-class prediction setting each class is associated with one set of regression coefficients, reflecting the impact of the predictors on the class prediction, and predictors are retained in the model when at least one associated regression coefficient is non-zero, irrespective from which class. Table 1 illustrates the regression coefficients sets obtained from a four-class classification setting with the lasso penalization scheme. The major disadvantage of the lasso penalization scheme relates to the selection of individual regression coefficients instead of complete predictors, resulting in the retention of an increased number of selected predictors.

The group lasso penalization scheme[8,9] overcomes this problem by defining a suitable penalization function. This penalization procedure could be considered as an intermediate between the lasso and ridge[10] penalization schemes and additionally has the attractive property of performing variable selection on predefined groups of predictors. Most logistic regression models, which have hitherto solely been based on single predictors, can now be replaced by entities reflecting group structures, e.g., pathways or gene sets. This predefined grouping has given the possibility to integrate prior knowledge into the model and create structures relevant to research; such as pathway analysis. The elastic net[11] was developed to take advantage of the grouping effect; however it lacks the ability to predefine group structures, which could inherently increase the interpretability of the derived prediction signature.

Table 1. Regression coefficients derived from a 4-class classification problem.

Probe set	Other	t(15;17)	t(8;21)	inv(16)	Gene symbol
1553588_at	9.55E-05	0	0	-0.0003	*ND3*
200026_at	9.91E-05	0	0	0	*RPL34*
200665_s_at	0	0	0	0.000659	*SPARC*
201324_at	-0.00024	0	0	0	*EMP1*
201360_at	-0.00014	0	0	0.000246	*CST3*
201432_at	0.00173	0	0	-0.00039	*CAT*
201502_s_at	0.000318	0	0	0	*NFKBIA*
201721_s_at	0	0	-0.00053	0	*LAPTM5*
202746_at	0	0	0	0.000388	*ITM2A*
202859_x_at	0	0	0.000122	0	*IL8*
202902_s_at	0	0	0	0.000201	*CTSS*
202917_s_at	0	0	0	0.00021	*S100A8*
203535_at	0	0	0	0.000762	*S100A9*

The lasso penalization scheme sets many individual regression coefficients to zero.

We have extended the group lasso penalization scheme for the multinomial logistic regression model. The imposition of a novel group structure results in the retention of complete predictors instead of individual regression coefficients, implying that retained predictors comprise non-zero regression coefficients for all defined classes. Interestingly, regression coefficients from the same predictor, belonging to the prediction signature of different classes, are now comparable. We utilize the gene expression data from a large cohort of AML patients (n=531), with distinct molecular entities adaptable as classification subjects, and demonstrate that the new penalization scheme has a prediction accuracy comparable to the lasso penalization scheme, with the retention of less predictors. We devised two classification problems to test our novel classification framework: (1) the AML entities harboring the favorable cytogenetic abnormalities inv(16)(p13q22), t(8;21)(q22;q22) or t(15;17)(q22;q12) and the mutually exclusive mutations in the gene *CEBPA*, (2) AML cases harboring combinations of mutations in *NPM1* or *FLT3* (internal tandem duplications, *FLT3*-ITD). In the former classification setting we demonstrate similar prediction efficiencies, compared to the lasso penalization scheme, with less predictors and in the latter case we demonstrate increased prediction efficiency, compared to previous efforts[12], with less predictors.

METHODS

Multinomial logistic regression

The multinomial logistic regression model is a multi-class prediction procedure, which predicts the probability of a class by fitting the data to a logistic curve. Initially, we have a specific number of observations; *n* (e.g., AML cases), a specific number of predictors; *p* (e.g., genes), and each

observation can be assigned to *g* outcome categories (e.g., classes). The outcome variables $Y_1,\cdots,Y_n$ are associated to each observation and a $n\times p$ matrix X, containing the data (e.g., gene expression levels), is constructed. For convenience the outcome variables are encoded by indicator functions corresponding to class participation. We define $y_{is} = \mathbf{1}_{\{Yi=s\}}$ ($i = 1\cdots, n; s = 1,\cdots, g$), and each class has its own regression coefficients vector, $\boldsymbol{\beta}_i \in \mathbb{R}^p (i = 1,\cdots,g)$. The corresponding probability is given by:

$$P(Y_i = s) = \mu_{is} = \frac{e^{\boldsymbol{\beta_{0s}} + \boldsymbol{x_i'\beta_s}}}{\sum_{t=1}^{\boldsymbol{g}} e^{\boldsymbol{\beta_{0t}} + \boldsymbol{x_i'\beta_t}}} (2)$$

The model defined in (2) is overparameterized as replacing (β_{k1}, ⋯, β_{kg}) by (β_{k1} + *c*, ⋯, β_{kg} + *c*), for any $k \in \{1,\cdots,p\}$ and $c \in \mathbb{R}$, results in the same probabilities. Commonly, this problem is solved by defining one outcome category as a reference category. The choice of reference category facilitates the interpretation of the resulting parameter estimates. Instead of choosing a reference category, we will treat the outcome categories as symmetrical[13,14] as penalized models are not invariant to setting reference categories resulting in different prediction rules. Furthermore, the penalized general linear models are not affected by overparameterization in terms of function identifiability problems. For notational convenience we rewrite the regression coefficient vectors into a long vector format: $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1,\cdots, \boldsymbol{\beta}_g)$. We also rewrite y_{is}, μ_{is} into $ng\times 1$ vectors: $\boldsymbol{y} = (y_{11},\cdots, y_{n1},\cdots, y_{1g},\cdots, y_{ng})$, $\boldsymbol{\mu} = (\mu_{11},\cdots,\mu_{n1},\cdots,\mu_{1g},\cdots,\mu_{ng})$ and the design matrix into $\boldsymbol{X} = X\otimes I_g$, where ⊗ is the Kronecker product. The log-likelihood of this model is:

$$\ell(\boldsymbol{\beta}^*) = \sum_{i=1}^{n}\sum_{s=1}^{g} y_{is}\, log(\mu_{is}) (3)$$

which has the gradient $\frac{\partial l(\beta)}{\partial \beta} = \boldsymbol{X'}(\boldsymbol{y} - \boldsymbol{\mu})$ and the Hessian $\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^r} = \boldsymbol{X'WX}$.
The $ng\times ng$ matrix W is given by:

$$\boldsymbol{W} = \begin{bmatrix} W^{11} & W^{12} & \cdots & W^{1g} \\ W^{21} & W^{22} & & \vdots \\ \vdots & & \ddots & \\ W^{g1} & \cdots & & W^{gg} \end{bmatrix}$$

Where

$$\operatorname{diag}(W^{st}) = \operatorname{diag}(W^{ts}) = \begin{cases} (-\mu_{1s}\mu_{1t},\cdots,-\mu_{ns}\mu_{nt})' & \text{if } s \neq t \\ \big(\mu_{1s}(1-\mu_{1s}),\cdots,\mu_{ns}(1-\mu_{ns})\big)' & \text{if } s = t \end{cases}$$

The Newton-Raphson algorithm is used to maximize the likelihood due to the convex likelihood function. Overparameterization of the model results into a singular Hessian matrix. Moore-Penrose or projection procedures resolve this issue, however, this caveat plays no role in the group lasso penalization scheme as it remains unaffected by overparameterization.

Penalty structure

The penalized log-likelihood under Lasso regulation (equation 1), imposes a L_1 regularization on each individual regression coefficient per predictor. Most regression coefficients become zero under strong penalization, resulting in sparse prediction rules. In essence the lasso penalization scheme selects individual regression coefficients rather than complete predictors, resulting in larger number of retained predictors than necessary. In addition, most regression coefficients become zero prohibiting the comparison of these coefficients between predefined classes. The group lasso penalization scheme allows for the definition of groups of predictors as entities of the model, instead of single regression coefficients, retaining predefined groups, facilitating interpretation of the obtained prediction signature. We propose a novel group lasso scheme for integration into the multinomial logistic regression model.[8,9] The beta matrix comprises columns of regression coefficient vectors affiliated to each class:

$$\tilde{\beta} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1g} \\ \beta_{21} & \beta_{22} & & \beta_{2g} \\ \vdots & & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pg} \end{bmatrix} = [\boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2 \quad \cdots \quad \boldsymbol{\beta}_g]$$

From the beta matrix many different group structures can be defined and illustrates the underlying mechanism of the group lasso penalization scheme. We would like to discard complete predictors; i.e. rows of the beta matrix, by setting all regression coefficients of the predictor simultaneously to zero. This is accomplished by defining each row vector of regression coefficients as a group in the group lasso penalization scheme. We have a *p*-dimensional feature vector $\boldsymbol{x}_i = \in \mathbb{R}^p$, which consists out of *J* groups and denote by df_j the degrees of freedom of group *j*, rewrite $\boldsymbol{x}_i = (\boldsymbol{x}'_{i,1}, \boldsymbol{x}'_{i,2}, \cdots, \boldsymbol{x}'_{i,J})$ and denote the group of variables by $\boldsymbol{x}'_{i,j} \in \mathbb{R}^{df_j}$, $j = 1, \cdots, J$. The regression coefficient vector is parameterized as $\boldsymbol{\beta}_t = (\beta_{0t}, \boldsymbol{\beta}_{1t}, \boldsymbol{\beta}_{2,t}, \cdots, \boldsymbol{\beta}_{J,t})'$, $t = 1, \cdots, G$.

Given the defined groups of regression coefficients we rewrite (equation 2) as:

$$P(Y_i = s) = \mu_{is} = \frac{e^{\beta_{0s} + \sum_{j=1}^{J} \boldsymbol{x}'_{i,j} \boldsymbol{\beta}_{j,s}}}{\sum_{t=1}^{\boldsymbol{g}} e^{\beta_{0t} + \sum_{j=1}^{J} \boldsymbol{x}'_{i,j} \boldsymbol{\beta}_{j,t}}} (4)$$

The group lasso estimator β_λ is given by the maximizer of the function:

$$\ell_{glasso}(\boldsymbol{\beta}^*; \lambda) = \ell(\boldsymbol{\beta}^*) - \lambda \sum_{j=1}^{J} \left\| \boldsymbol{\beta}_j \right\|_2 = \ell(\boldsymbol{\beta}^*) - \psi(\boldsymbol{\beta}^*) \ (5)$$

Hence, the penalty function sums the norm of each row of the beta matrix $\tilde{\beta}$. Note, we integrate the square root of the degrees of freedom of each group in the summation, as described previously.[8,9] The degrees of freedom term is omitted due to equal size for all groups.

Group lasso estimator

To optimize the penalized log-likelihood function (equation 5), the low-memory BFGS algorithm (L-BFGS-B)[15] is used. This Quasi-Newton algorithm necessitates a limited number of previous function and gradient evaluations to estimate the inverse Hessian. The gradient of the penalized log-likelihood function is given by:

$$\frac{\partial \ell_{glasso}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = \frac{\partial \ell(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} - \lambda \frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} \quad (6)$$

where the gradient of the penalty function is defined as:

$$\frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \beta_{ij}} = \frac{\partial}{\partial \beta_{ij}} \left(\sqrt{\beta_{11}^2 + \cdots + \beta_{1g}^2} + \cdots + \sqrt{\beta_{p1}^2 + \cdots + \beta_{pg}^2} \right) = \frac{\beta_{ij}}{\sqrt{\beta_{i1}^2 + \cdots + \beta_{ig}^2}} \quad (7)$$

Reparameterization and parameter identifiability

Optimizing the penalized log-likelihood function leads to major problems, as the function is only strictly convex and continuous in all internal subspaces of the regression coefficients. The derivative of the penalized log-likelihood function remains undefined when one of the regression coefficients equals zero. This issue is resolved by reparameterizing the model to a higher dimension where the function is strictly convex and continuous. The following reparameterization is proposed:

$$\beta_{ij} = \beta_{ij}^+ - \beta_{ij}^- \begin{cases} \beta_{ij}^+ = \max(\beta_{ij}, 0), \beta_{ij}^+ \geq 0 \\ \beta_{ij}^- = -\min(\beta_{ij}, 0), \beta_{ij}^- \geq 0 \end{cases}$$

The reparameterization is realized by decomposing the individual regression coefficients into a positive part function (PPF) and a negative part function (NPF). These functions are constrained by the fact that each must be non-negative. For this reason we make use of the box constraints definable in the L-BFGS-B algorithm. Note that at the convergence either the PPF, NPF or both should be equal to zero. This reparameterization results in a model with twice as many parameters, which are restricted to a subspace of non-negative regression coefficients. In this single subspace the penalized log-likelihood function is strictly convex, continuous, and is differentiable at each internal point. Hence, instead of dealing with distinct continuous subspaces where the function is non-differentiable at their borders, i.e. when one of the regression coefficients is set to zero, we now have one subspace where the function is differentiable. The log-likelihood gradient remains unchanged under the reparameterization, but the penalty function gradients are given by:

$$\frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \beta_{ij}^+} = \frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \beta_{ij}} \frac{\partial \beta_{ij}}{\partial \beta_{ij}^+} = \frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \beta_{ij}}$$

$$\frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \beta_{ij}^-} = \frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \beta_{ij}} \frac{\partial \beta_{ij}}{\partial \beta_{ij}^-} = -\frac{\partial \psi(\boldsymbol{\beta}^*)}{\partial \beta_{ij}}$$

A problem occurs when all the regression coefficients of a group become zero, as the penalty function is no longer differentiable. To solve this problem the following limit is taken for the sake of continuity:

$$\lim_{\beta_{ij}\to 0} \frac{\beta_{ij}}{\sqrt{\beta_{i1}^2 + \cdots + \beta_{ig}^2}} = 1, \;\; if\; \beta_{i1} = \cdots = \beta_{i(j-1)} = \beta_{i(j+1)} = \cdots = \beta_{ig} = 0$$

Reparameterization affects the optimization of the penalized log-likelihood due to the parameter identifiability problem. The penalty function $\psi(\boldsymbol{\beta}^*)$ (equations 5-7) sums the norms of the row vectors of the beta matrix $\widetilde{\beta}$, determined by the squared regression coefficients β_{ij}^2 belonging to the same group. Under the reparameterization this squared regression coefficient is given by:

$$\beta_{ij}^2 = (\beta_{ij}^+ - \beta_{ij}^-)^2 = \beta_{ij}^{+2} - 2\beta_{ij}^+\beta_{ij}^- + \beta_{ij}^{-2} \quad (8)$$

Multiple instances of β_{ij}^+ or β_{ij}^- could give the exact same β_{ij}^2 (equation 8). This problem can be resolved by imposing a constraint on this equation. At convergence either the PPF, NPF or both should be equal to zero. This implies that the middle term of the factorization of β_{ij}^2 should be forced to be zero. This leads to the following redefinition of equation 8:

$$\beta_{ij}^2 = \beta_{ij}^{+2} + \beta_{ij}^{-2} \quad (9)$$

Redefinition of the penalty function leads to the following rewritten penalized log-likelihood function:

$$\ell_{glasso}(\boldsymbol{\beta}^*; \lambda) = \ell(\boldsymbol{\beta}^*) - \lambda \sum_{j=1}^{J} \sqrt{\left\|\boldsymbol{\beta}_j^+\right\|_2^2 + \left\|\boldsymbol{\beta}_j^-\right\|_2^2} \quad (10)$$

The triangle-inequality shows that:

$$\sqrt{\left\|\boldsymbol{\beta}_j^+\right\|_2^2 + \left\|\boldsymbol{\beta}_j^-\right\|_2^2} \geq \sqrt{\left\|\boldsymbol{\beta}_j^+ - \boldsymbol{\beta}_j^-\right\|_2^2} = \sqrt{\left\|\boldsymbol{\beta}_j^+\right\|_2^2 - 2\left(\boldsymbol{\beta}_j^+\right)^T\boldsymbol{\beta}_j^- + \left\|\boldsymbol{\beta}_j^-\right\|_2^2} \quad (11)$$

Hence, the redefined penalty function $\psi(\boldsymbol{\beta}^*)$ is always larger or equal than its original definition. Given the inequality (equation 11) and the fact that either the PPF, NPF or both are zero at convergence, the redefined penalty function becomes equal to the original definition. By this redefinition we have solved the parameter identifiability problem and proven the obtainment of the exact same prediction rules without convergence problems.

Table 2 illustrates the results from the same 4-class classification problem, defined earlier, based on the modified group lasso penalization scheme. In comparison with Table 1 it becomes clear that: (1) the number of predictors is decreased (2) none of the regression coefficient of the retained predictors became zero, and (3) the new group structure facilitates comparison of the regression coefficients between classes.

Table 2. Regression coefficients from a 4-class classification problem with the modified group lasso.

Probe set	Other	t(15;17)	t(8;21)	inv(16)	Gene symbol
1553588_at	0.00018085	-8.59E-05	5.97E-05	-0.0001546	*ND3*
200665_s_at	-0.00014592	-1.73E-05	-7.23E-05	0.000235584	*SPARC*
201324_at	-0.00017254	1.18E-05	2.64E-05	0.000134331	*EMP1*
201360_at	-0.00020149	9.67E-06	-6.17E-05	0.000253532	*CST3*
201432_at	0.000946723	-0.00025838	-3.35E-05	-0.0006548	*CAT*
201502_s_at	0.000131047	-0.00012284	6.43E-05	-7.25E-05	*NFKBIA*
201721_s_at	0.000325746	1.74E-05	-0.00034902	5.93E-06	*LAPTM5*
202746_at	-0.00012466	1.67E-05	-0.00011551	0.000223436	*ITM2A*
202902_s_at	-7.06E-06	-1.00E-05	-5.58E-06	2.27E-05	*CTSS*
202917_s_at	-6.06E-05	-0.0001884	-8.16E-05	0.000330612	*S100A8*
203535_at	-0.00018007	-6.25E-05	-9.28E-05	0.000335433	*S100A9*

The modified group lasso penalization scheme produces sparser prediction rules and facilitates the comparison of regression coefficients between classes.

RESULTS

AML is not a single disease, but a group of neoplasms with various genetic aberrations and variable prognosis and response to treatment.[16,17] The search for novel molecular markers is essential for therapeutical decision-making. A large number of molecular markers have been identified in the last decade, however, the complete underlying mechanism of leukomogenesis still remains elusive. With the use of gene expression profiling (GEP), the challenge lies in generating reliable prediction rules that can discriminate the different AML subtypes; for instance for the improvement of treatment decisions or classification. We applied our algorithm to the GEPs of 540 clinically and molecularly well-characterized AML cases originating from two different cohorts. The first cohort comprises 269 AML cases previously analyzed[12,18], while the second cohort was subsequently generated and analyzed.[12,19,20] All samples are analyzed with Affymetrix Human Genome U133 Plus 2.0 GeneChips (Affymetrix, Santa Clara, CA, USA). All clinical, cytogenetic and molecular information as well as the gene expression data are available at Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo, accession number GSE6981). All data has been preprocessed as previously described.[12] AML cohort 1 (n=269) has been used as training set while AML cohort 2 (n=261) is subsequently used as test set. The optimal value for the regularization parameter λ was determined by 5-fold cross-validation. The gene expression signatures are available in Supplementary Tables 1 and 2.

Classification problem 1: AML entities with favorable cytogenetics or mutations in *CEBPA*

Background and classification objective

The first classification objective concerns the correct classification of AML cases belonging to favorable risk categories, i.e. with the cytogenetic abnormalities inv(16)(p13q22), t(8;21)(q22;q22) or t(15;17)(q22;q12). An additional class was created for AML cases harboring mutations in the transcription factor *CEBPA*[19-21] associated with myelopoiesis.[22] Usually, these molecular abnormalities are mutually exclusive, and no overlapping gene expression patterns were expected. Finally, an additional class 'Other', was created for the remaining AML cases. Table 3 depicts the distribution of the different classes over the two cohorts.

Table 3. **Distribution of the AML samples over the predefined classes.**

Classes	AML cohort 1 (n=261)	AML cohort 2 (n=264)	Risk
Other	180 (70%)	204(77%)	-
t(15;17)	18(7%)	7(3%)	Favorable
t(8;21)	22(8%)	16(6%)	Favorable
inv(16)	23(8%)	18(7%)	Favorable
CEBPa	18(7%)	18(7%)	Favorable

Results

Initially, we applied the global test for the multinomial logistic regression model[23] to test whether the GEPs contain any information for the discrimination of the AML subtypes. This hypothesis test determines whether the global expression patterns significantly relate to the AML subtypes. The H_0-hypothesis was rejected ($p < 0.0001$) implying that the GEPs have discriminatory power. The optimal regularization parameter λ for the modified group lasso penalization scheme was estimated to be 50 by 5-fold cross-validation, resulting in a predictive signature comprising 74 probe sets (Supplementary Table 1). Figure 1 illustrates the estimated test error curve for eleven evaluations of λ. The optimal regularization parameter λ for the lasso penalization was determined by the same cross-validation procedure. The regularization parameter was set at 0.02 with 75 retained probe sets (Supplementary Table 3). For this classification problem it does not matter whether to select the lasso or the modified group lasso penalization scheme with respect to the number of retained predictors. The retained predictors of both procedures greatly overlap, with the exception of a few predictors. Strikingly, the lasso makes four additional miss-classifications compared to the group lasso (Table 5 vs. Supplementary Table 6).

AML subtypes harboring the prognostic favorable cytogenetic abnormalities (t(15;17), t(8;21) and inv(16)) were predicted with 100% accuracy (Table 5), which was consistent with previous work.[12] A substantial proportion of the samples with *CEBPA* mutations were classified as belonging to the 'Other' category. After further investigation it became apparent that all misclassified

samples contain a single mutation in *CEBPA* (*CEBPA*^{sm}) instead of biallelic mutations (*CEBPA*^{dm}). Previous work noted that double, but not single mutated samples have a distinct GEP and can be accurately predicted.[19,20] Furthermore, overall survival (OS) is significantly different between the *CEBPA*^{sm} and *CEBPA*^{dm} cases (Figure 2), indicating that *CEBPA*^{dm} AML cases have a more favorable prognosis, compared to *CEBPA*^{sm} and *CEBPA*^{wt} AML cases, but also a distinct gene expression signature.

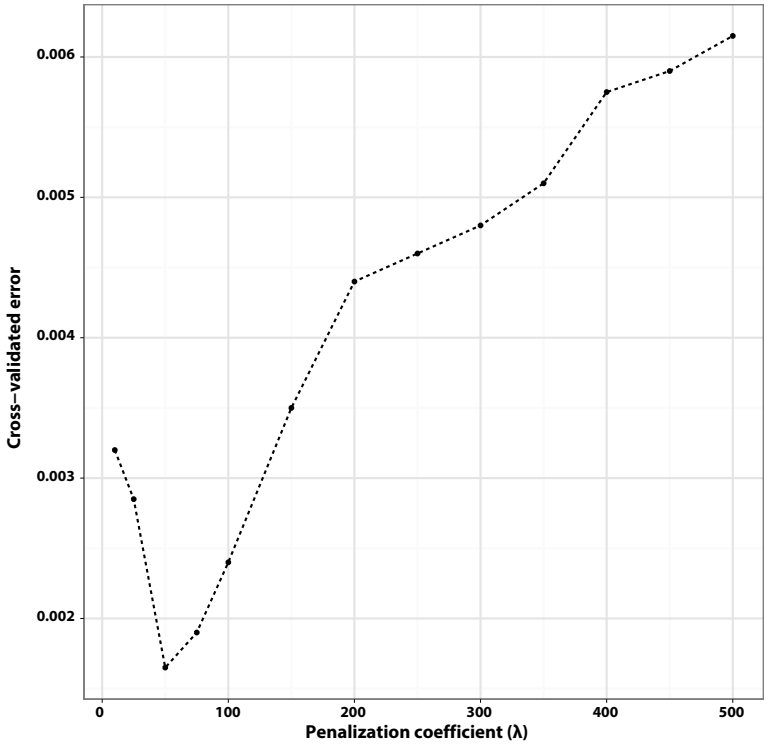


Figure 1. 5-fold cross-validated error curve for classification problem 1.

Interpretation

In addition to determining the prediction accuracy, the interpretation of the obtained prediction signatures could provide vital insight into AML pathobiology. Previous work by Kohlmann et al.[24] demonstrated accurate discrimination of the same set of AML subtypes (excluding *CEBPA* mutations) by selecting 23 genes. We extracted the gene expression levels of 17 genes belonging to the 74-gene prediction signature and performed clustering of the genes (Figure 3 [Top]). The bottom of Figure 3 contains regression coefficients for a subset of the genes taken from the obtained prediction rule, demonstrating that the regression coefficients of each gene strongly reflect the up- or downregulation tendency of that specific class.

Note that the retained predictors are not always fully explanatory for the underlying leukemogenic mechanism. For instance for inv(16) AML cases, the partial inversion of chromosome 16 results in a fusion protein, namely CBFB-MYH11. Due to the fusion, the expression level of *MYH11* is substantially increased compared to other AML subtypes and is used as a gene expression marker for this particular AML subtype. Many classification algorithms based on differential expression would select this gene, however, this is not the case when the lasso or group lasso penalization schemes are applied. These penalization schemes select only one predictor if there is a group of pair-wise correlated genes and this could be the case for *MYH11*.

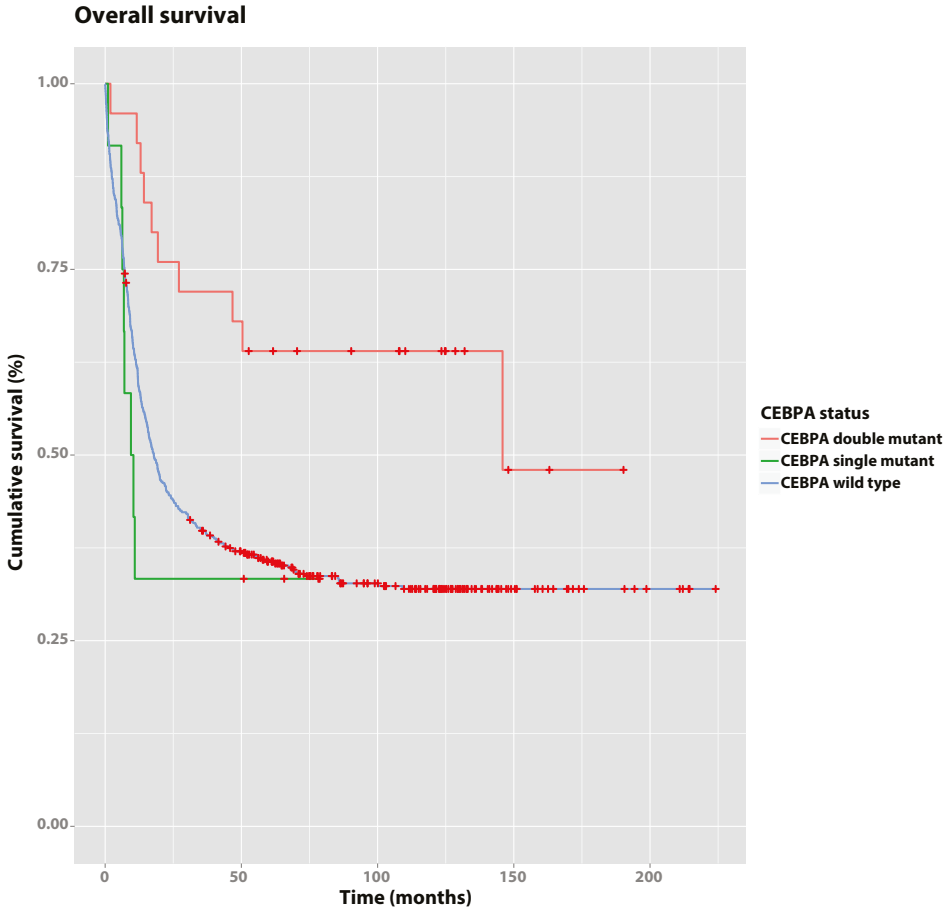
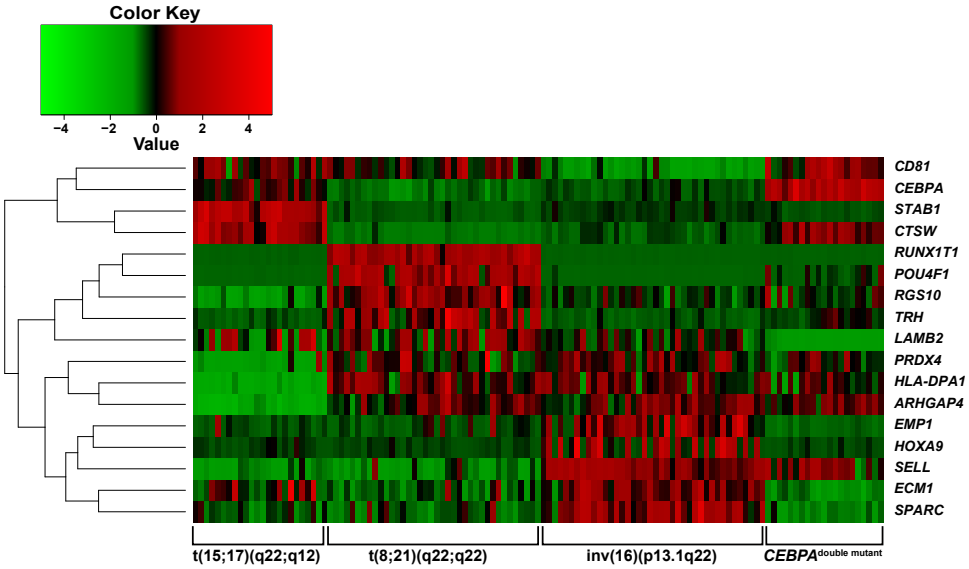


Figure 2. Overall survival and event-free survival. Overall survival among CEBPA^{dm}, CEBPA^{sm} and CEBPA^{wt} AML cases, pooled: p=0.011.



Probe set	Other	t(15;17)	t(8;21)	inv(16)	*CEBPA*^{mut}	Gene
200665_s_at	-6.62E-06	-4.68E-07	-3.00E-06	1.08E-05	-6.94E-07	*SPARC*
200675_at	0.000442	8.25E-07	-3.19E-06	-0.00019	-0.00025	*CD81*
204039_at	-0.00096	0.000151	-0.00062	-0.00068	0.002107	*CEBPA*
204150_at	-7.40E-05	0.000173	-6.08E-05	-5.21E-05	1.41E-05	*STAB1*
204563_at	-0.00015	-0.00011	-0.00013	0.000199	0.000194	*SELL*
205529_s_at	-0.0001	-4.91E-05	0.000257	-9.36E-05	-7.36E-05	*RUNX1T1*
206940_s_at	-4.49E-05	-2.83E-05	0.000163	-7.21E-05	-1.71E-05	*POU4F1*
211990_at	-0.00012	-0.00022	0.000108	-1.26E-05	0.000243	*HLA-DPA1*

Figure 3. (Top) Clustering. Clustered genes, colors of the cells relate to up- or downregulation of the gene for that particular sample: (green) downregulation, (red) upregulation. (Bottom) Regression coefficients for a selection of genes.

Previous work by Wunderlich et al.[25] has demonstrated that other genes, such as *SPARC* and *EMP1*, are highly correlated with *MYH11* in inv(16) AML cases. Figure 4 illustrates that these genes are significantly upregulated in inv(16) AML cases. Additionally, these genes belong to the top 20 of highest upregulated genes in inv(16) compared to other groups (Supplementary Table 5). In conclusion, we demonstrated that the imposed group structure on the beta matrix $\tilde{\beta}$ results in less retained predictors with equal prediction accuracy compared to the lasso penalization scheme. In addition, many retained genes (Supplementary Table 1) have been previously associated with leukemogenesis. For example, the genes *HOXA9* and *TRIB1* are known to be deregulated in AML, and have been identified as cooperative genes together with *MEIS1*.[26,27]

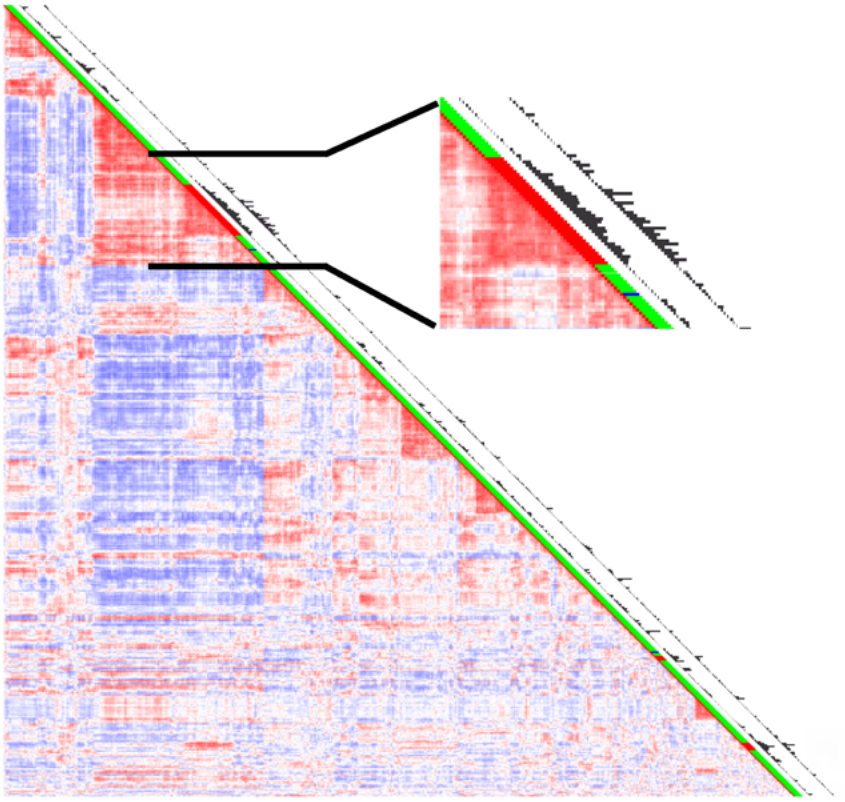


Figure 4. Gene expression levels of *SPARC* and *EMP1* are elevated in the inv(16) AML subgroup. Correlation view of the 531 AML patients. Colors of the cells relate to pair-wise Pearson's correlation coefficient values: red indicates higher positive and blue indicates higher negative correlation between samples. The inv(16) aberration status is indicated by the first row next to the heatmap (red, mutant; green, wild-type). Histograms next to the heatmap reflect the expression levels *SPARC* and *EMP1* respectively, and shows a significant elevated expression for the inv(16) AML cases.

Classification problem 2: *NPM1* mutations and *FLT3*-ITD

Background and classification objective

The *NPM1* mutation is the most frequent molecular abnormality observed in AML.[28] NPM1 is predominantly found in the nucleolus and is thought to be an important chaperone protein and to play a role in ribosome biogenesis.[29] NPM1 is disrupted by mutations introducing a nuclear export signal, which replaces the nucleolar localization signal, resulting in its displacement to the cytoplasm.[30] It has been demonstrated that *NPM1* mutations frequently coincide with *FLT3*-ITDs and frequently occur in AML patients with a normal karyotype. The *NPM1* mutation is considered a favorable prognostic marker with respect to overall and event-free survival.[28]

Table 4. Distribution of the AML samples over the predefined classes.

Classes	AML cohort 1 (n=261)	AML cohort 2 (n=268)
NPM1-/*FLT3*-ITD-	149(57%)	160(60%)
NPM1+/*FLT3*-ITD-	44(17%)	32(12%)
NPM1-/*FLT3*-ITD+	28(11%)	33(12%)
NPM1+/*FLT3*-ITD+	40(15%)	43(16%)

FLT3 is a receptor tyrosine kinase protein situated on the cell membrane, where it is activated by cytokines[31] initiating a cascade of transduction signals through secondary messengers, such as STAT5[32], and is known to play an important role in cell differentiation, survival and proliferation. Frequently the gene *FLT3* contains an internal tandem duplication which contributes to the development of AML. Furthermore, the *FLT3*-ITD aberration is considered a poor prognostic marker with respect to overall and event-free survival.

In this classification problem, we classify patients which have the *NPM1* mutation alone (*NPM1*+/*FLT3*-ITD-), *FLT3*-ITD alone (*NPM1*-/*FLT3*-ITD+), both mutations (*NPM1*+/*FLT3*-ITD+) or none (NPM1-/*FLT3*-ITD-). Table 4 depicts the distribution of these classes.

Table 5. Prediction accuracy established with the group lasso penalization scheme integrated into the multinomial logistic regression model.

Classes		Test set error		Sensitivity	Specificity	Predictive Value	
		Neg	Pos	%	%	Neg	Pos
Case 1							
	Other	6/81	0/180	100	93	100	97
	t(15;17)	0/243	0/18	100	100	100	100
	t(8;21)	0/239	0/22	100	100	100	100
	inv(16)	0/238	0/23	100	100	100	100
	CEBPA^{mut}	0/243	6/18	67	100	98	100
Case 2							
	Other	23/119	7/160	96	81	93	87
	NPM1+/*FLT3*-ITD-	17/237	9/32	72	93	96	58
	NPM1-/*FLT3*-ITD+	6/236	23/33	30	97	91	63
	NPM1+/*FLT3*-ITD+	10/226	17/43	60	96	93	72

The following calculations were used for evaluation: sensitivity=true positives/(true positive + false negatives), specificity=true negatives/(true negatives + false positives), positive predictive value=true positives/(true positives + false positives), negative predictive value=true negatives /(true negatives + false negatives)

Results

The global test determined that the GEPs contain discriminatory power given the defined AML subtypes ($p < 0.0001$). With 5-fold cross-validation, we have determined the optimal regularization parameter (λ=375) for the group lasso penalization scheme as illustrated in Figure 5. The model

retained 110 probe sets (Supplementary Table 2). For the lasso penalization scheme we determined the optimal regularization parameter to be 10 with 152 retained probe sets. The group lasso penalization scheme substantially decreased the number of necessary predictors with a similar prediction accuracy compared to the lasso penalization scheme. The group lasso penalization scheme misclassifies 56 AML cases whereas the lasso penalization scheme misclassifies 62 AML cases (Table 5 vs. Supplementary Table 6).

Previous gene expression analyses have demonstrated that *NPM1* mutations are strongly associated with a discriminative *HOX*-signature.[28,32] The obtained prediction signature indicates that the *HOXA9* and *HOXB3* gene expression levels have a strong impact on the classification of *NPM1*+ AML cases. A relatively high number of AML cases were misclassified as having the *NPM1* mutation. This could have several reasons: (1) many false positives contained an 11q23 abnormality affecting the MLL protein, which is an important *HOX* gene expression regulator[28] (2) some *FLT3*-ITD AML cases also exhibit strong *HOX* gene expression deregulation.

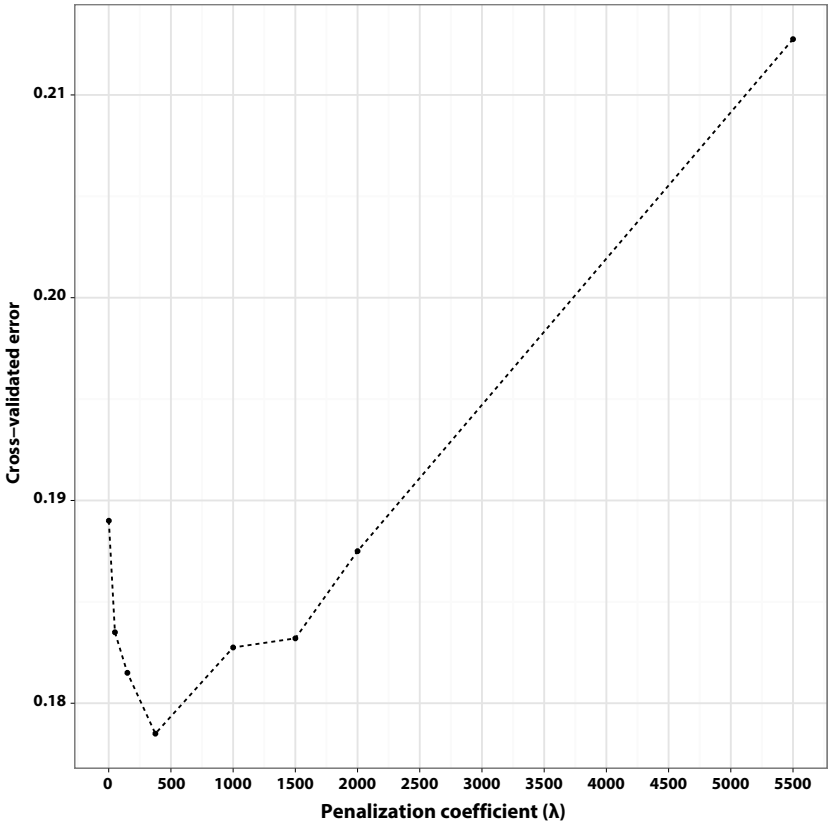


Figure 5. 5-fold cross-validated error curve for classification problem 2.

The major classification problem stem from AML cases harboring the *FLT3*-ITD abnormality. AML cases harboring the abnormality can only be moderately classified as indicated in Table 5, possibly due to the following reasons: (1) the *FLT3*-ITD abnormality is sub-clonal. AML cases with a sub-clonal *FLT3*-ITD abnormality might have a weak discriminative expression signature (2) a subgroup of AML cases harboring the *FLT3*-ITD abnormality have different gene expression patterns due to concurrent mutations.

We can conclude that determining samples with a *FLT3*-ITD abnormality is difficult based on their GEP alone. Most of the *NPM1*-/*FLT3*-ITD+ samples are misclassified as wild-type (*NPM1*-/*FLT3*-ITD-). This is in line with the observation that some, if not most, of the *NPM1*-/*FLT3*-ITD+ AML cases exhibit a weak distinctive GEP. The same holds for the *NPM1*+/*FLT3*-ITD+ samples which are mostly misclassified as *NPM1*+/*FLT3*-ITD-, and vice versa. It seems that the lack of a discriminative *FLT3*-ITD expression signature makes it difficult to concurrently predict all classes with a high accuracy.

3.2.3 Interpretation

The retained predictors show an enrichment for ribosomal, heatshock, immunoglobulin and *HOX* genes. Many genes in the gene expression signature are related to processes of cellular stress, inflammation response and DNA repair mechanisms. The large number of ribosomal genes present in the signature could be due: (1) the activation of DNA repair or cell homeostasis mechanisms in response to stress conferred by the molecular abnormalities, (2) the *NPM1* mutation results in the dislocation of the protein from the nucleolus to the cytoplasm. The protein is known as a chaperon protein for ribosomes; however these results could indicate that it may also be involved in ribosome biogenesis.

DISCUSSION

The aim of this study was to develop a sparse multi-class classification model based on the group lasso penalization scheme in the multinomial logistic regression model. To create such an algorithm, we have developed a new group structure based on the beta matrix. This group structure facilitates the selection of an entire predictor instead of individual regression coefficients. We have demonstrated that the prediction accuracy is similar to that of the lasso penalization scheme, yet with the retention of less predictors. To illustrate that our approach is effective we have applied the algorithm on microarray gene expression data of a large cohort of well characterized AML patients. Not only have we demonstrated that the group lasso penalization scheme achieves good prediction accuracy, but also that it obtains a sparse prediction rule containing many previously affiliated genes. We would like to note that the algorithm does not always converge under specific circumstances, such as situations with very low number of cases, which might be related to numerical instability.

We have demonstrated that our algorithm behaves as expected and we would like to make note that many different group structures can be developed. We expect in the near future that singular entities in contemporary classification procedures will be replaced by group structures, which increase the interpretability of the prediction signature and generate the opportunity to analyze different aspects of the model. As a final remark we would like to conclude that the development of novel group structures could increase the interpretability of the prediction rule, the prediction accuracy, and possibly further our understanding of cancer and its pathogenesis.

ACKNOWLEDGEMENTS

We are indebted to Peter J.M. Valk, Justine K. Peeters, Erdogan Taskesen (Erasmus Medical Center, Rotterdam, The Netherlands) who provided us with data and biologic technical support. We are grateful to Marcel J.T. Reinders (Technical University of Delft, Delft, The Netherlands) who provided us with helpful feedback.

REFERENCES

1. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001;96(456):1348-1360.
2. Krishnapuram B, Carin L, Figueiredo MA, Hartemink AJ. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2005;27(6):957-968.
3. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 2010;33(1):1.
4. Cawley GC, Talbot NL, Girolami M. Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in neural information processing systems*. 2007;19:209.
5. Le Cessie S, Van Houwelingen J. Ridge estimators in logistic regression. *Applied statistics*. 1992:191-201.
6. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996:267-288.
7. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics*. 2004;5(3):427-443.
8. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006;68(1):49-67.
9. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008;70(1):53-71.
10. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55-67.
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005;67(2):301-320.
12. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *haematologica*. 2009;94(1):131-134.
13. Goeman JJ, le Cessie S. A goodness-of-fit test for multinomial logistic regression. *Biometrics*. 2006;62(4):980-985.
14. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biom J*. 2010;52(1):70-84.
15. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Mathematical programming*. 1989;45(1-3):503-528.
16. TCGA. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.
17. Sanders MA, Valk PJ. The evolving molecular genetic landscape in acute myeloid leukaemia. *Curr Opin Hematol*. 2013;20(2):79-85.
18. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *New England Journal of Medicine*. 2004;350(16):1617-1628.
19. Taskesen E, Bullinger L, Corbacioglu A, 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*. 2011;117(8):2469-2475.
20. Wouters BJ, Löwenberg B, Erpelinck-Verschueren CA, van Putten WL, Valk PJ, Delwel R. 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*. 2009;113(13):3088-3091.
21. Pabst T, Eyholzer M, Haefliger S, Schardt J, Mueller BU. Somatic CEBPA mutations are a frequent second event in families with germline CEBPA mutations and familial acute myeloid leukemia. *J Clin Oncol*. 2008;26(31):5088-5093.

22. Zhang P, Iwasaki-Arai J, Iwasaki H, et al. Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBP alpha. *Immunity*. 2004;21(6):853-863.

23. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004;20(1):93-99.

24. Kohlmann A, Schoch C, Schnittger S, et al. Molecular characterization of acute leukemias by use of microarray technology. *Genes Chromosomes Cancer*. 2003;37(4):396-405.

25. Wunderlich M, Krejci O, Wei J, Mulloy JC. Human CD34+ cells expressing the inv (16) fusion protein exhibit a myelomonocytic phenotype with greatly enhanced proliferative ability. *Blood*. 2006;108(5):1690-1697.

26. Röthlisberger B, Heizmann M, Bargetzi MJ, Huber AR. TRIB1 overexpression in acute myeloid leukemia. *Cancer genetics and cytogenetics*. 2007;176(1):58-60.

27. Jin G, Yamazaki Y, Takuwa M, et al. Trib1 and Evi1 cooperate with Hoxa and Meis1 in myeloid leukemogenesis. *Blood*. 2007;109(9):3998-4005.

28. Verhaak RG, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood*. 2005;106(12):3747-3754.

29. Lindstrom MS. NPM1/B23: A Multifunctional Chaperone in Ribosome Biogenesis and Chromatin Remodeling. *Biochem Res Int*. 2011;2011:195209.

30. Falini B, Nicoletti I, Bolli N, et al. Translocations and mutations involving the nucleophosmin (NPM1) gene in lymphomas and leukemias. *haematologica*. 2007;92(4):519-532.

31. Levis M, Small D. FLT3: ITDoes matter in leukemia. *Leukemia*. 2003;17(9):1738-1752.

32. Alcalay M, Tiacci E, Bergomas R, et al. Acute myeloid leukemia bearing cytoplasmic nucleophosmin (NPMc+ AML) shows a distinct gene expression profile characterized by up-regulation of genes involved in stem-cell maintenance. *Blood*. 2005;106(3):899-902.

CHAPTER 3

Prognostic impact, concurrent genetic mutations and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML: further evidence for CEBPA double-mutant AML as a distinctive disease entity

Erdogan Taskesen[1], Lars Bullinger[2], Andrea Corbacioglu[2], Mathijs A. Sanders[1], Claudia A. Erpelinck-Verschueren[1], Bas J. Wouters[1], Sonja C. van der Poel-van de Luytgaarde[1], Frederik Damm[3], Jürgen Krauter[3], Arnold Ganser[3], Richard F. Schlenk[2], Bob Löwenberg[1], Ruud Delwel[1], Hartmut Döhner[2], Peter J.M. Valk[1,*], and Konstanze Döhner[2,*].

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands

[2] University of Ulm, Department of Internal Medicine III, Ulm, Germany

[3] Hannover Medical School, Department of Hematology, Hemostasis, Oncology, and Stem Cell Transplantation, Hannover, Germany

* These authors contributed equally to this work

Blood. 2011 Feb 24;117(8):2469-75

ABSTRACT

We evaluated concurrent gene mutations, clinical outcome, and gene expression signatures of *CEBPA* double (*CEBPA*^{dm}) versus single (*CEBPA*^{sm}) mutations in 1182 cytogenetically normal AML (CN-AML) patients (16-60 years). We identified 151 (12.8%) patients with *CEBPA* mutations (91 *CEBPA*^{dm} and 60 *CEBPA*^{sm}). The incidence of germline mutations was 7% (5 out of 71), including three C-terminal mutations. *CEBPA*^{dm} patients had a lower frequency of concurrent mutations than *CEBPA*^{sm} patients (P<.0001). Both, *CEBPA*^{dm} and *CEBPA*^{sm} were associated with favorable outcome compared to *CEBPA*^{wt} [5-year overall survival (OS), 63% and 56% versus 39%; P<.0001 and P=.05, respectively]. However, in multivariable analysis only *CEBPA*^{dm} was a prognostic factor for favorable outcome [OS, hazard ratio (HR): .36, P<.0001; event-free survival, HR: .41, P<.0001; relapse-free survival, HR: .55, P=.001)]. Outcome in *CEBPA*^{sm} is dominated by concurrent *NPM1* and/or *FLT3* internal tandem duplication (ITD) mutations. Unsupervised and supervised GEP analyses showed that *CEBPA*^{dm} AML (n=42), but not *CEBPA*^{sm} AML (n=18) expressed a unique gene signature. A 25-probeset prediction signature for *CEBPA*^{dm} AML showed 100% sensitivity and specificity. Based on these findings, we propose that *CEBPA*^{dm} should be clearly defined from *CEBPA*^{sm} AML and considered as a separate entity in the classification of AML.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter3/

INTRODUCTION

In the current World Health Organization (WHO) classification of acute myeloid leukemia (AML), "AML with mutated *CEBPA* (CCAAT/enhancer binding protein alpha)" has been designated as a provisional disease entity in the category "AML with recurrent genetic abnormalities".[1,2] *CEBPA* encodes a transcription factor that is essential for neutrophil development. Targeted disruption of *Cebpa* in mice results in a selective block in early granulocyte development, a hallmark of AML.[3,4] Two proteins may be translated from the *CEBPA* mRNA transcript, i.e., a 42kDa (p42) and a shorter 30kDa (p30) protein. The p42 isoform contains two regulatory transactivation domains (TAD1 and TAD2) in the N-terminus, while the shorter p30 isoform only carries the TAD2 domain. Both isoforms contain the C-terminal basic DNA-binding domain and the leucine zipper (bZIP), involved in DNA-binding and protein dimerization. In AML, *CEBPA* mutations mainly occur in cytogenetically normal AML (CN-AML) with an incidence of 5-14%.[5-14] Two main types of mutations can be distinguished: N-terminal frame-shift mutations resulting in the translation of the 30-kDa protein only, and the C-terminal in-frame mutations in the basic zipper region affecting DNA-binding and homo- and heterodimerization.[8,15] As a consequence, these mutations create an imbalance between proliferation and differentiation of hematopoietic progenitors.[10,16]

AML with *CEBPA* mutations can be separated into two subgroups, i.e., those with a single mutation (*CEBPA*^{sm}) and those with double mutations (*CEBPA*^{dm}).[17-21] In the majority of *CEBPA*^{dm} AML, both alleles are mutated.[19] These biallelic mutations frequently consist of an N-terminal mutation on one allele and a C-terminal bZIP mutation on the other. In *CEBPA*^{sm} AML, mutations occur either in the N- or in the C-terminus of the gene. In previous studies, in which these two subgroups were not considered, AML with mutated *CEBPA* had a relatively good outcome.[5,7,12,13] More recent data suggest that this favorable outcome is mainly observed in AML with *CEBPA*^{dm} and not *CEBPA*^{sm}.[17-21] Moreover, it has been suggested that concurrent mutations may occur more frequently in *CEBPA*^{sm} than in *CEBPA*^{dm} AML. The impact of coexisting mutations remains elusive and needs to be validated in large cohorts.

By applying gene expression profiling (GEP), it was demonstrated that *CEBPA*^{dm} AML can be distinguished from *CEBPA*^{sm} and the majority of *CEBPA*^{wt} AML based on a characteristic signature.[18] However, a *CEBPA*^{dm} GEP signature did not predict *CEBPA*^{dm} AML with maximum accuracy, since AML in which *CEBPA* was silenced by promoter hypermethylation (*CEBPA*^{silenced}) carried a highly similar signature.[22,23] Objectives of this study were to evaluate the impact of *CEBPA*^{dm} versus *CEBPA*^{sm} on clinical outcome of CN-AML and to investigate the impact of concurrent *NPM1*^{mutant} and/or *FLT3*^{ITD}. In addition, we searched for *CEBPA*-associated gene signatures and determined the frequency of predisposing *CEBPA* germline mutations. For these purposes, we combined data sets from the Dutch-Belgian Hemato-Oncology Cooperative Group (HOVON) and Swiss Group for Clinical Cancer Research (SAKK) and the German-Austrian AML Study Group (AMLSG).

PATIENTS AND METHODS

Patients and molecular analyses

Diagnostic bone marrow (BM) or peripheral blood (PB) samples from 1182 younger adults (16-60 years) with CN-AML were analyzed; 193 patients were enrolled on HOVON/SAKK protocols -04, -04A, -29, and -42 (available at www.hovon.nl)[24-27], and 989 patients on AMLSG protocols AMLHD93 (n=74)[28], AML HD98A (n=313)[29], AMLSG 07-04 (n=376; ClinicalTrials.gov Identifier NCT00151242), AML SHG 02-95 (n=94)[30], and AML SHG 01-99 (n=180, ClinicalTrials.gov Identifier NCT00209833). All patients provided written informed consent in accordance with the Declaration of Helsinki. All trials were approved by the Institutional Review Board of Erasmus University Medical Center, University of Ulm, and Hannover Medical School.

Mutation analyses for the genes *FLT3* (internal tandem duplications [ITD] and tyrosine kinase domain mutations [TKD]) and *NPM1* were performed as described previously.[31-33] *CEBPA*^{dm} and *CEBPA*^{sm} AML were identified by denaturing high-performance liquid chromatography (dHPLC) or direct sequencing as described.[18] Cases that carried an insertion polymorphism[18,21] (http://www.ncbi.nlm.nih.gov/sites/snp; http://genome.ucsc.edu/cgi-bin/hgGateway; http://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene) or variation(s) that did not lead to amino acid changes were considered wild-type. Cases were categorized as *CEBPA*^{dm} when two different mutations or one homozygous mutation were present as determined by sequencing analysis; cases with only a single heterozygous mutation were designated as *CEBPA*^{sm}. In 71 of the 151 patients with *CEBPA* mutations, DNA obtained from buccal swabs (n=52), PB (n=8) or BM (n=11) in complete remission (CR) was studied for the presence of *CEBPA* germline mutations. Patient demographics and molecular characteristics are summarized in Table 1. All *CEBPA*-mutated patients, except for 07-04 treated patients within the AMLSG protocol, have been previously reported in different studies.[7,13,18]

Gene expression profiling

Data from GEP analysis were available for 674 AML (53% CN-AML, HOVON-SAKK and AMLSG-cohorts), generated using Affymetrix (Santa Clara, CA, USA; Table S1). Sample processing and quality control were carried out as described previously.[23,34] For both cohorts, normalization of raw data was processed with Affymetrix Microarray Suite 5 (MAS5) to target intensity values at 100. Intensity values were log2 transformed and mean centered per probeset per cohort. GEP data are available at the NCBI Gene Expression Omnibus [accession numbers GSE14468 (HOVON-SAKK cohort) and GSE22845 (AMLSG-cohort)]. There were 42 *CEBPA*^{dm} and 18 *CEBPA*^{sm} cases for which the GEP was determined (Table S1).

Table 1. Demographics, clinical and molecular characteristics of CEBPA^{wt}, CEBPA^{sm}, and CEBPA^{dm} CN-AML.

Characteristic	*CEBPA*^{wt}	*CEBPA*^{sm}	*P*, *CEBPA*^{sm}	*CEBPA*^{dm}	*P*, *CEBPA*^{dm}	*P*, *CEBPA*^{sm}
	(n = 1031)	(n = 60)	vs *CEBPA*^{wt}	(n = 91)	vs *CEBPA*^{wt}	vs *CEBPA*^{dm}
Median age, years (range)	48 (16-60)	46 (18-60)	0.28	44 (16-60)	0.04*	0.66
Sex, n (%)			0.79		0.74	0.74
Male	500 (48)	28 (47)		46 (51)		
Female	531 (52)	32 (53)		45 (49)		
WBC count, x10^9/L			0.23		0.062	0.86
Median (range)	28 (0.2-372)	25 (1.1-345)		28 (1.5-262)		
Missing	34	1		4		
Platelet count, x10^9/L			0.77		< 0.0001*	< 0.0001*
Median (range)	65 (5-746)	62 (10-361)		38 (4-265)		
Missing	40	3		4		
Bone marrow blasts			0.83		0.53	0.76
Median (range)	80 (0-100)	80 (0-97)		78 (2-100)		
Missing	80	7		4		
Molecular abnormalities						
FLT3^{ITD}, n (%)	347 (33.7)	18 (30)	1	7 (7.7)	< 0.0001*	0.00015*
Missing	69	9		5		
FLT3^{TKD}, n (%)	95 (9.2)	4 (6.7)	0.81	2 (2.2)	0.018*	0.2
Missing	48	6		3		
NPM1$^+$, n (%)	560 (54.3)	21 (35)	0.018*	3 (3.3)	0	< 0.0001*
Missing	88	10		8		

Number of cases (percentage), median, range, or missing values are indicated. WBC indicates white blood cell. *P < .05 computed using the Mann-Whitney U test (continuous variables) and 2-sided Fisher exact test (categorical variables).

Statistical analyses

Statistical analyses were performed using Mathworks (Matlab R2009b) with the statistical, bioinformatics and pattern recognition toolbox (Prtools). For clinical, molecular, univariate and multivariate analyses, patients with CN-AML and age ≤ 60 (Table S1) were included. Molecular and clinical variables of both patient cohorts (HOVON-SAKK and AMLSG) were comparable. Differences were assessed for *CEBPA*^{sm} and *CEBPA*^{dm} groups in comparison with *CEBPA*^{wt} group (Table 1), by using the Mann-Whitney-*U* test for continuous variables and the two-sided Fisher exact test for categorical variables.

For univariate analysis significance was assessed using the stratified log-rank test and Kaplan-Meier estimates for overall survival (OS), event-free survival (EFS) and relapse-free survival (RFS). The recommended consensus criteria[35] were used for the definition of CR and survival. Multivariate analysis was performed by using a stratified Cox's proportional hazard model. For all analyses, a *P*-value ≤ .05 was considered statistically significant and for survival analyses, *P*-values

were computed using the full time span. Note that the closed testing procedure[36] was applied and a correction for multiple testing[37] was only performed if the global log-rank test resulted in a *P*-value ≤ .05. For gene expression-based classification of *CEBPA*^{dm} cases, GEP of the HOVON-SAKK cohort was used to derive the 25-probeset predictive signature and the AMLSG-cohort as validation set. To summarize, a logistic regression model with Lasso regularization (a continuous feature selection procedure) was used as it takes the correlation structure between the probesets into account (see Supplementary material).

RESULTS

Frequency and types of acquired *CEBPA*^{dm} and *CEBPA*^{sm} mutations

CEBPA mutations were detected in 151 of the 1182 (12.8%) CN-AML; 91 (60%) patients had *CEBPA*^{dm}, within these the combination of N- and C-terminal mutations was the predominant genotype (82/91). *CEBPA*^{dm} cases with only N-terminal or C-terminal mutations were less frequently observed (4/91 and 5/91, respectively). Sixty of the 151 (40%) *CEBPA*-mutated cases had *CEBPA*^{sm} which occurred most frequently in the N-terminus (47/60). Only 13 of the 60 *CEBPA*^{sm} cases had in-frame insertion or deletion mutations affecting the bZIP domain (Figure 1).

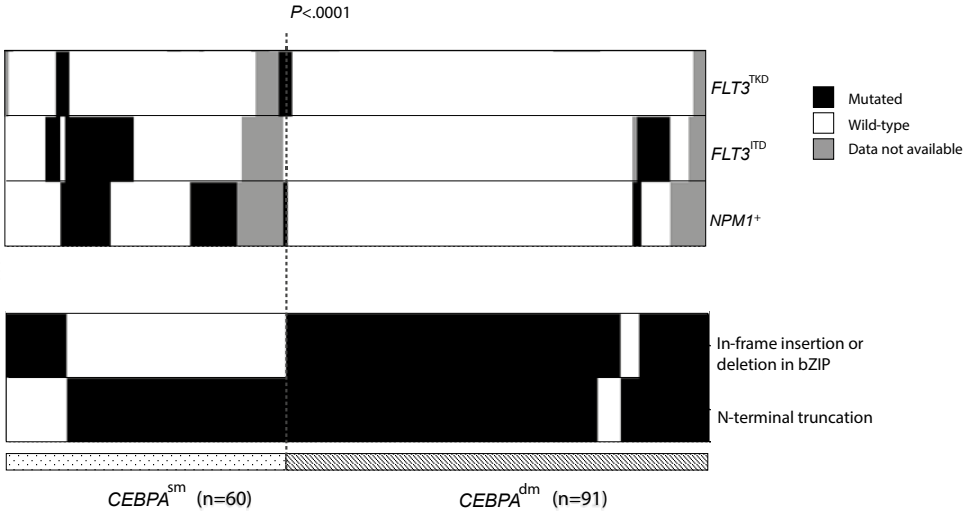


Figure 1. Distribution of concurrent mutations in *CEBPA*^{dm} and *CEBPA*^{sm} patients. Columns represent patients and rows the genotypes *FLT3*^{TKD}, *FLT3*^{ITD} and *NPM1*^{mutant} (black), wildtype (white) or missing (grey). The in-frame insertion or deletion in bZIP and N-terminal truncation mutations in *CEBPA* are highlighted in black.

Table 2. Germline patient demographics and molecular characteristics.

Patient ID	Age at diagnosis, (year)	Germline mutation	Acquired mutation*	Additional mutation†	Familial AML	History *CEBPA* mutation
98A-751	28	338delC	1080insGAA	None	Yes	*CEBPA*^{dm}
07/04-268 (ULM_10)	25	307delG	1122_1123ins 1075_1225	*KRAS*, *WT1*	Yes	*CEBPA*^{dm}
BioID 769	51	1096T>C	478_485del	None	No	*CEBPA*^{dm}
98A-543	33	1164G>A	None	*FLT3*^{TKD}, *NPM1*	No	*CEBPA*^{sm}
07/04-48 (ULM_20)	59	1036G>T	1086insAAG	None	No	*CEBPA*^{dm}

Characteristics of 5 of 71 (7%) *CEBPA*-mutant AML patients who carried *CEBPA* germline mutations. *KRAS*, Kirsten Rat sarcoma; *WT1*, Wilms tumor 1. * Data according to GenBank accession no. Y11525.

Association of acquired CEBPA^{dm} and *CEBPA*^{sm} mutations with concurrent gene mutations and clinical characteristics

Concurrent mutations were seen less frequently in *CEBPA*^{dm} than in *CEBPA*^{sm} AML (22% versus 60%; *P*<.0001, Figure 1); frequencies for *NPM1*^{mutant} were 3.3% and 35% (*P*<.0001), and for *FLT3*^{ITD} were 7.7% and 30% (*P*=.00015), respectively (Table 1). When comparing *CEBPA*^{sm} and *CEBPA*^{wt} AML, *NPM1*^{mutant} were slightly less frequent in *CEBPA*^{sm} AML (35% versus 54.3%; *P*=.018); the frequency of *FLT3*^{ITD} was comparable between the two groups (30% versus 33.7%). Regarding presenting clinical characteristics, *CEBPA*^{dm} mutations were associated with younger age (median 44 versus 48 years; *P*=.04) and lower platelet counts (median 38x10^9/L versus 65x10^9/L; *P*<.0001) compared with *CEBPA*^{wt} patients (Table 1).

Impact of CEBPA^{dm} and *CEBPA*^{sm} on response to induction therapy and clinical outcome

For clinical outcome analyses, 1182 CN-AML were considered. *CEBPA*^{dm} was associated with a higher CR rate when compared with *CEBPA*^{sm} (92% versus 78%, *P*=.02) and *CEBPA*^{wt} (92% versus 79%, *P*=.002). There was no difference in CR probability between *CEBPA*^{sm} and *CEBPA*^{wt} patients (78% versus 79%, *P*=.86). The median follow-up time for survival in the 1182 CN-AML patients was 33 months (95%-CI, 25.6 to 40.4); the estimated 5-year OS and RFS were 42% (95%-CI, 39% to 45%) and 34% (95%-CI, 31% to 38%), respectively.

CEBPA^{dm} AML was associated with a significantly superior outcome compared with *CEBPA*^{wt} AML (5-year OS, 63% versus 39%, *P*<.0001; EFS, 45% versus 28%, *P*<.0001; RFS, 44% versus 32%, *P*=.05; Figures 2A and Supplementary Figures S3A and S3D). A somewhat better outcome was also found for *CEBPA*^{sm} AML compared with *CEBPA*^{wt} AML (5-year OS, 55% versus 39%, *P*=.05; RFS, 49% versus 32%, *P*=.02; but not EFS, 37% versus 28%, *P*=.22). No significant difference was evident between *CEBPA*^{dm} and *CEBPA*^{sm} AML (5-year OS, *P*=.06; EFS, *P*=.16; RFS, *P*=.48). Of note, no differences in outcome were observed between *CEBPA*^{sm} patients with either C-terminal (n=13)

or N-terminal (n=47) mutations (5-year OS, 54% versus 56%, *P*= .58; Figure S4). In multivariate analyses considering other prognostic indicators (listed in Table 3), the presence of *CEBPA*^{dm} was an independent prognostic factor for favorable OS (HR, .36, *P*<.0001), EFS (HR, .41, *P*<.0001) and RFS (HR, .55, *P*=.001), whereas *CEBPA*^{sm} did not impact these three endpoints (Table 3).

Table 3. Multivariate analysis for overall, event-free and relapse-free survival in CN-AML.

Variables	HR	95% CI	*P*-value
Overall survival			
CEBPA$^{sm\alpha}$	0.70	0.46 - 1.07	0.1
CEBPA$^{dm\alpha}$	0.36	0.23 - 0.55	< 0.0001*
FLT3$^{ITD\beta}$	1.78	1.49 - 2.14	< 0.0001*
FLT3$^{TKD\beta}$	0.84	0.61 - 1.15	0.28
NPM1$^{+\beta}$	0.56	0.46 - 0.67	< 0.0001*
WBC count$^{\delta}$, x10^9/L	1.35	1.12 - 1.62	< 0.0001*
Age$^{\epsilon}$	1.02	1.01 - 1.03	< 0.0001*
Event-free survival			
CEBPA$^{sm\alpha}$	0.86	0.6 - 1.22	0.4
CEBPA$^{dm\alpha}$	0.41	0.29 - 0.57	< 0.0001*
FLT3$^{ITD\beta}$	1.56	1.33 - 1.84	< 0.0001*
FLT3$^{TKD\beta}$	0.8	0.6 - 1.07	0.13
NPM1$^{+\beta}$	0.45	0.39 - 0.53	< 0.0001*
WBC count$^{\delta}$, x10^9/L	1.27	1.08 - 1.5	0.003*
Age$^{\epsilon}$	1.01	1.01 - 1.02	0.003*
Relapse-free survival			
CEBPA$^{sm\alpha}$	0.79	0.51 - 1.22	0.3
CEBPA$^{dm\alpha}$	0.55	0.38 - 0.79	0.001*
FLT3$^{ITD\beta}$	1.75	1.45 - 2.12	< 0.0001*
FLT3$^{TKD\beta}$	0.82	0.59 - 1.13	0.22
NPM1$^{+\beta}$	0.56	0.46 - 0.68	< 0.0001*
WBC count$^{\delta}$, x10^9/L	1.33	1.1 - 1.61	0.002*
Age$^{\epsilon}$	1.01	1 - 1.02	0.001*

Stratified Cox's proportional hazard model for multivariable analysis of *CEBPA*^{dm} and *CEBPA*^{sm} as prognostic markers for overall, event-free and relapse-free survival. Analyses included 1182 CN-AML patients with age ≤ 60. Subgroup: $^{\alpha}$ *CEBPA* status versus *CEBPA*^{wt}, $^{\beta}$ *FLT3*^{ITD} versus no *FLT3*^{ITD} mutation, $^{\beta}$ *FLT3*^{TKD} versus no *FLT3*^{TKD} mutation, $^{\beta}$ *NPM1*^{mutant} versus no *NPM1*^{wt}, $^{\delta}$ WBC count higher than 20x10^9/L versus lower than 20x10^9/L, $^{\epsilon}$ Age is used as continuous variable
**P*-value ≤ 0.05

Treatment outcome of AML with *CEBPA*^{sm} is dominated by *FLT3/NPM1* genotypes

We performed explorative subgroup analyses in *CEBPA*^{sm} and *CEBPA*^{wt} AML to evaluate the impact of four *FLT3*/*NPM1* genotype subgroups: *FLT3*^{ITD}/*NPM1*^{mutant} (n=10); *FLT3*^{ITD}/*NPM1*^{wt} (n=8); *FLT3*^{wt}/*NPM1*^{mutant} (n=11); and *FLT3*^{wt}/*NPM1*^{wt} (n=21). Ten cases from the *CEBPA*^{sm} group were excluded

for which the genotypes were unknown. Among patients with *CEBPA*^{sm} AML, the *FLT3*^{ITD}/*NPM1*^{wt} genotype had an inferior OS compared to those with the *FLT3*^{wt}/*NPM1*^{wt} genotype (5-year OS, 25% versus 49%, *P*=.05; Figure 2B); for EFS and RFS, there was a trend towards an inferior outcome (Figure S3B and S3E); in contrast, the *FLT3*^{wt}/*NPM1*^{mutant} genotype associated in trend with a favorable outcome compared with the *FLT3*^{wt}/*NPM1*^{wt} genotype (5-year OS, 78% versus 49%, *P*=.2, EFS: 59% versus 32%, *P*=.08, RFS: 66% versus 40%, *P*=.38, Figure 2B, S3B and S3E). In analogy, in the *CEBPA*^{wt} group the *FLT3*^{ITD}/*NPM1*^{wt} genotype had a significantly inferior survival compared with the *FLT3*^{wt}/*NPM1*^{wt} genotype (5-year OS, 17% versus 34%, *P*=.001; EFS, 11% versus 14%, *P*=.04; RFS, 15% versus 24%, *P*=.002; Figure 2C, S3C and S3F), whereas the *FLT3*^{wt}/*NPM1*^{mutant} genotype was associated with a favorable outcome (5-year OS, 57% versus 34%, *P*<.0001; EFS, 47% versus 14%, *P*<.0001; RFS: 50% versus 24%, *P*<.0001; Figure 2C, S3C and S3F). Thus, we observed comparable trends for favorable (*FLT3*^{wt}/*NPM1*^{mutant}) and inferior (*FLT3*^{ITD}/*NPM1*^{wt}) outcome in the *CEBPA*^{sm} and *CEBPA*^{wt} subgroups. The outcome for all *CEBPA*^{sm} *FLT3*/*NPM1* genotypes was higher (not significant, *P*> .05), compared to the *CEBPA*^{wt} genotypes, however, the distinct groups were relatively small. For *CEBPA*^{dm} AML, sample sizes of the composite genotypic subgroups were too small for analysis.

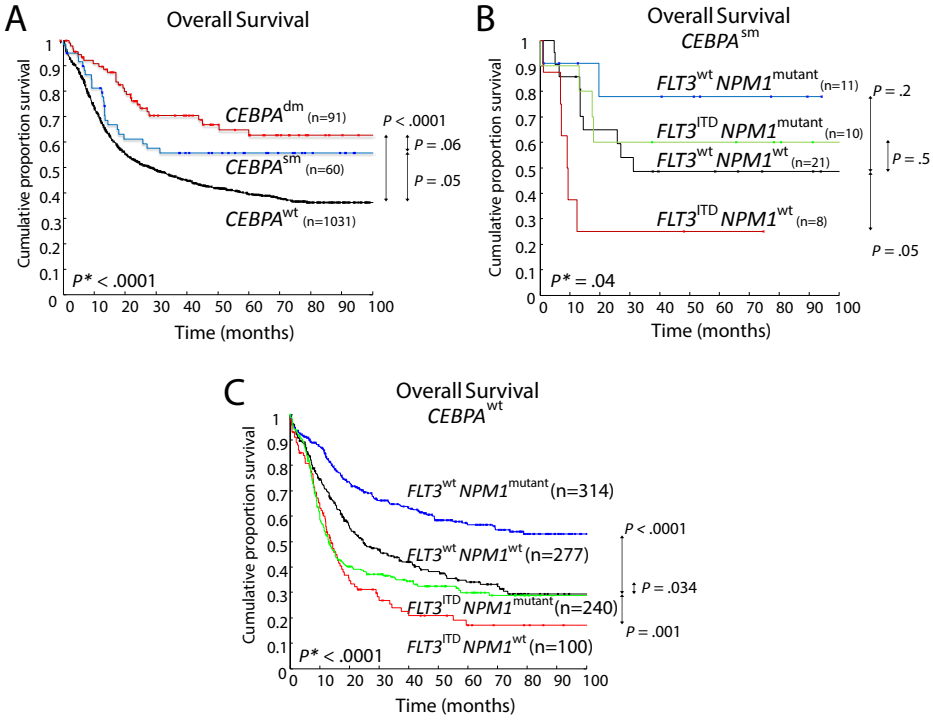


Figure 2. Kaplan-Meier survival curves of overall survival. (A) Kaplan-Meier survival curves for OS among the three groups *CEBPA*^{dm}, *CEBPA*^{sm} and *CEBPA*^{wt}. (B) Survival curves for OS of the four genotypes *FLT3*^{ITD}/*NPM1*^{mutant}, *FLT3*^{ITD}/*NPM1*^{wt}, FLT3wt/*NPM1*^{mutant}, and *FLT3*^{wt}/*NPM1*^{wt} within the *CEBPA*^{sm} group. (C) Survival curves for OS of the four genotypes within *CEBPA*^{wt}. * *P*-value by the global log-rank test.

Unsupervised analyses of GEP showed homogeneity in *CEBPA*^{dm} AML cases

GEP was performed in a subset of the CN-AMLs patients and also includes cytogenetically abnormal patients (Table S1; n=674). Unsupervised analyses, by hierarchical clustering, revealed distinct GEP clusters (Figure 3A), including the known clusters of AML with inv(16), t(15;17) or t(8;21), as shown previously.[23] These subtypes revealed high correlation within the GEP cluster (average correlation: .42, .49 and .49, respectively) and differed significantly between the AML cases with any of these aberrations (*P*<.0001, *P*<.0001, and *P*<.0001, Figure S5B, S5C and S5E). We observed that the *CEBPA*^{dm} AML cases were highly similar within the cluster (average correlation: .35) and differed significantly from cases without a *CEBPA*^{dm} (*P*<.0001, Figure S5D). *CEBPA*^{sm} AML cases showed reduced similarity (average correlation: .15) and did not differ from cases without *CEBPA*^{sm} (*P*=.12, Figure S5A and Figure 3A).

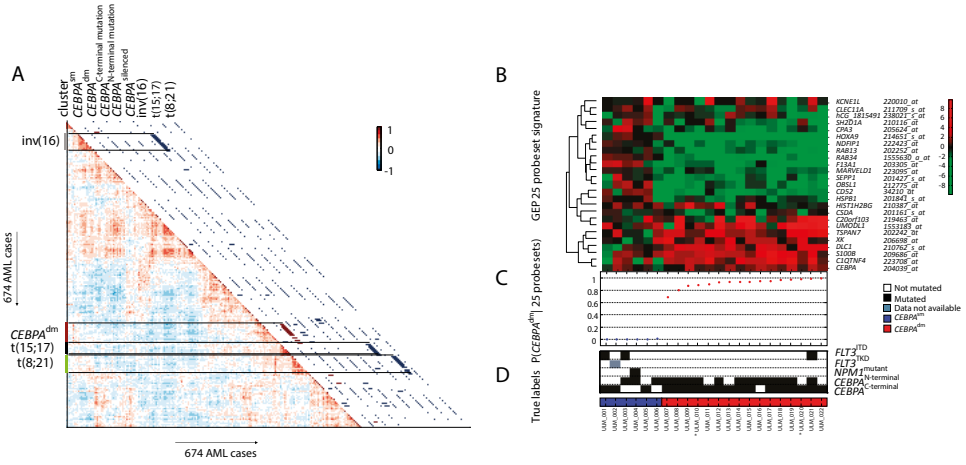


Figure 3. Unsupervised analyses and classification results of candidate *CEBPA*^{dm} cases with their gene expression profile and their molecular characteristics. (A) Pair-wise correlations between the 674 AML cases; color coding cells: positive (red) or negative (blue) correlations, as indicated by the scale bar. Presence of molecular abnormalities, indicated at the top of the plot, are depicted on the diagonal by a red or blue bar. *CEBPA*$^{C\text{-terminal mutation}}$ and *CEBPA*$^{N\text{-terminal mutation}}$ indicates the presence of homozygous mutations. (B) Gene expression levels of *CEBPA*^{dm} and *CEBPA*^{sm} patients for the defined 25-probeset signature. The colors are relative to the mean per probeset. (C) Ordered computed posterior probabilities for the *CEBPA*^{dm} signature given the 25-probeset signature. (D) (black) presence of mutation, (white) wildtype, (grey) missing value. * germline *CEBPA*^{dm} cases.

CEBPA^{dm} AML is accurately predicted based on GEP

The previously predictive *CEBPA*^{dm} signature[18] was hampered by the recently reported *CEBPA* silenced AML cases that carry a similar GEP.[22] Two independent AML cohorts were used to train and evaluate the predictive value of the *CEBPA*^{dm} signature in terms of sensitivity and specificity.

A predictive signature was created, containing 25-probesets by using a logistic regression model with Lasso regularization (Figure 3B and Supplementary material Table S2)[38,39], which selects discriminative probesets between the classes, *CEBPA*^{dm} (n=26) and all other AML cases, i.e., *CEBPA*^{wt} and *CEBPA*^{sm} (n=494). Subsequently, a classifier was trained on the entire HOVON-SAKK cohort based on a two class approach; 26 *CEBPA*^{dm} versus 494 cases (*CEBPA*^{wt} and *CEBPA*^{sm}). This trained classifier subsequently classified 16 candidate *CEBPA*^{dm} cases (Supplementary Table S3) in the AMLSG-cohort out of 154 AML cases (16 *CEBPA*^{dm}, 6 *CEBPA*^{sm} and 132 *CEBPA*^{wt}: Supplementary material Table S1). Among the *CEBPA*^{dm} cases were 5 cases with either homozygous N- or C-terminal *CEBPA*^{dm} mutations, and a *CEBPA*^{dm} patient with a germline C-terminal mutation. This approach showed perfect sensitivity and specificity (both 100%, Figure 3C). In addition, we performed a classification between *CEBPA*^{dm}, *CEBPA*^{sm}, and *CEBPA*^{wt} to infer, if we were able to accurately classify *CEBPA*^{sm} cases. We observed that all *CEBPA*^{sm} cases were classified as *CEBPA*^{wt}, thus *CEBPA*^{sm} cases do not have a consistent gene expression pattern and were different from the *CEBPA*^{dm} group.

DISCUSSION

Here, we established the value of the *CEBPA*^{dm} mutation in a large cohort of CN-AML patients from AMLSG and HOVON-SAKK treatment trials. We detected 91 (7.7%) double *CEBPA* and 60 (5.1%) single *CEBPA* mutations among 1182 patients. In multivariate analyses, we demonstrate that the presence of *CEBPA*^{dm}, but not *CEBPA*^{sm}, is an independent factor for favorable outcome in AML, which confirms previous findings reported in studies with relatively small cohorts.[17-19,21]

Concurrent mutations were significantly less frequent in *CEBPA*^{dm} compared with *CEBPA*^{sm} AML. This was true for *FLT3*^{ITD} and in particular for *NPM1*^{mutant} that were virtually not present among *CEBPA*^{dm} cases, a finding that is consistent with previously published data.[20] Compared to previous studies[17-21], and the large number of cases, we were able to evaluate the prognostic impact of the *CEBPA* mutational status in the context of the *FLT3*/*NPM1* genotypes. Among *CEBPA*^{sm} AML, the four combined genotypes showed similar trend with regard to outcome as compared with *CEBPA*^{wt} AML (Figure 2B and 2C). Nevertheless, we observed a higher outcome (not significant) for all *CEBPA*^{sm} *FLT3*/*NPM1* genotypes compared to the *CEBPA*^{wt} genotypes, but these groups are relatively small. These findings, supported by data from multivariable analysis, strongly suggest that not the existence of *CEBPA*^{sm} per se but rather the combined effects of *CEBPA*^{sm} and *FLT3*^{ITD} and/or *NPM1*^{mutant} determine outcome in these AML patients.

Here, we generated a refined GEP signature, consisting out of 25-probesets that predict *CEBPA*^{dm} AML cases. This signature showed sensitivity and specificity of 100% and has a better predictive power than the *CEBPA*^{dm} signature previously defined.[18] In fact, in contrast to the previous signature, the new signature also discriminates *CEBPA*^{dm} from AML with hypermethylation of the proximal promoter region of *CEBPA*.[22] Classification results were not affected by homozygous N- or C-terminal *CEBPA*^{dm} mutations or those due to germline mutation. Currently, nucleotide

sequencing is used as the gold standard for the identification of *CEBPA* mutations. Due to the much higher effort of gene expression profiling this technique should not be considered as a primary diagnostic tool in AML. However, GEP can be confirmatory, especially in cases where the *CEBPA* gene appears difficult to sequence. More importantly, GEP provides relevant insights in the biology of the disease and the affected signaling pathways and therefore allows further classification/refinement of AML.

Finally, we evaluated the frequency of *CEBPA* germline mutations in this large cohort of *CEBPA*-mutated cases. Among 71 mutated patients, 5 revealed germline mutations. Of these cases 4 developed *CEBPA*^{dm} AML, i.e., these cases acquired a mutation in the second allele, in line with previous data.[40,41] Interestingly, we for the first time identified 3 C-terminal germline mutations of which 2 cases acquired a second *CEBPA* mutation at the time of AML diagnosis. In GEP analysis both cases clustered within the *CEBPA*^{dm} group and were classified as a *CEBPA*^{dm}, providing evidence that these C-terminal sequence variations are mutations rather than polymorphisms. All 3 C-terminal germline mutations were predicted to be damaging for the function and the structure of the protein.

In the current World Health Organization (WHO) AML classification, "AML with mutated *CEBPA*" has been designated as a provisional disease entity in the category "AML with recurrent genetic abnormalities". Based on our data obtained from a large patient cohort together with findings from previous reports we propose that *CEBPA*^{dm} AML should be clearly distinguished from *CEBPA*^{sm} AML and that only "AML with *CEBPA*^{dm}" should be considered as an independent entity in the classification of the disease.

Acknowledgments

The authors thank Martin van Vliet and Jelle Goeman for the discussions. This research was supported by the Center for Translational Molecular Medicine (CTMM) and supported by grants P38/05//A49/05//F03 [Else Kröner-Fresenius-Stiftung], 01GI9981 [Network of Competence Acute and Chronic Leukemias], 01KG0605 [IPD-Meta-Analysis: A model-based hierarchical prognostic system for adult patients with acute myeloid leukemia (AML)] from the Bundesministerium für Bildung und Forschung (BMBF), Germany.

Contribution

E.T. performed research, data analysis, data interpretation, and manuscript writing; L.B. and A.C. performed research, data analysis and interpretation; M.A.S. performed data analysis, data interpretation and manuscript writing; C.A.J.E., B.J.W., and S.C.P.L. performed research; F.D. performed research and data interpretation; J.K., and A.G. provided provision of study material; R.F.S. performed research, data interpretation and manuscript writing; B.L., R.D., H.D., P.J.M.V., and K.D. designed the study, performed data interpretation and manuscript writing.

REFERENCES

1. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood.* Jul 30 2009;114(5):937-951.
2. Dohner H, Estey EH, Amadori S, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood.* Jan 21;115(3):453-474.
3. Rosenbauer F, Tenen DG. Transcription factors in myeloid development: balancing differentiation with transformation. *Nat Rev Immunol.* Feb 2007;7(2):105-117.
4. Zhang DE, Zhang P, Wang ND, Hetherington CJ, Darlington GJ, Tenen DG. Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein alpha-deficient mice. *Proc Natl Acad Sci U S A.* Jan 21 1997;94(2):569-574.
5. Barjesteh van Waalwijk van Doorn-Khosrovani S, Erpelinck C, Meijer J, et al. Biallelic mutations in the CEBPA gene and low CEBPA expression levels as prognostic markers in intermediate-risk AML. *Hematol J.* 2003;4(1):31-40.
6. Bienz M, Ludwig M, Leibundgut EO, et al. Risk assessment in patients with acute myeloid leukemia and a normal karyotype. *Clin Cancer Res.* Feb 15 2005;11(4):1416-1424.
7. Frohling S, Schlenk RF, Stolze I, et al. CEBPA mutations in younger adults with acute myeloid leukemia and normal cytogenetics: prognostic relevance and analysis of cooperating mutations. *J Clin Oncol.* Feb 15 2004;22(4):624-633.
8. Gombart AF, Hofmann WK, Kawano S, et al. Mutations in the gene encoding the transcription factor CCAAT/enhancer binding protein alpha in myelodysplastic syndromes and acute myeloid leukemias. *Blood.* Feb 15 2002;99(4):1332-1340.
9. Mueller BU, Pabst T. C/EBPalpha and the pathophysiology of acute myeloid leukemia. *Curr Opin Hematol.* Jan 2006;13(1):7-14.
10. Nerlov C. C/EBPalpha mutations in acute myeloid leukaemias. *Nat Rev Cancer.* May 2004;4(5):394-400.
11. Pabst T, Mueller BU, Zhang P, et al. Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia. *Nat Genet.* Mar 2001;27(3):263-270.
12. Preudhomme C, Sagot C, Boissel N, et al. Favorable prognostic significance of CEBPA mutations in patients with de novo acute myeloid leukemia: a study from the Acute Leukemia French Association (ALFA). *Blood.* Oct 15 2002;100(8):2717-2723.
13. Schlenk RF, Dohner K, Krauter J, et al. Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. *N Engl J Med.* May 1 2008;358(18):1909-1918.
14. Snaddon J, Smith ML, Neat M, et al. Mutations of CEBPA in acute myeloid leukemia FAB types M1 and M2. *Genes Chromosomes Cancer.* May 2003;37(1):72-78.
15. Asou H, Gombart AF, Takeuchi S, et al. Establishment of the acute myeloid leukemia cell line Kasumi-6 from a patient with a dominant-negative mutation in the DNA-binding region of the C/EBPalpha gene. *Genes Chromosomes Cancer.* Feb 2003;36(2):167-174.
16. Calkhoven CF, Muller C, Leutz A. Translational control of C/EBPalpha and C/EBPbeta isoform expression. *Genes Dev.* Aug 1 2000;14(15):1920-1932.
17. Pabst T, Eyholzer M, Fos J, Mueller BU. Heterogeneity within AML with CEBPA mutations; only CEBPA double mutations, but not single CEBPA mutations are associated with favourable prognosis. *Br J Cancer.* Apr 21 2009;100(8):1343-1346.
18. Wouters BJ, Lowenberg B, Erpelinck-Verschueren CA, van Putten WL, Valk PJ, Delwel R. 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.* Mar 26 2009;113(13):3088-3091.
19. Dufour A, Schneider F, Metzeler KH, 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.* Feb 1;28(4):570-577.

20. Green CL, Koo KK, Hills RK, Burnett AK, Linch DC, Gale RE. 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.* May 3.

21. Hou HA, Lin LI, Chen CY, Tien HF. Reply to 'Heterogeneity within AML with CEBPA mutations; only CEBPA double mutations, but not single CEBPA mutations are associated with favorable prognosis'. *Br J Cancer.* Aug 18 2009;101(4):738-740.

22. Figueroa ME, Wouters BJ, Skrabanek L, et al. Genome wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood.* Mar 19 2009;113(12):2795-2804.

23. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene expression profiles in acute myeloid leukemia. *N Engl J Med.* Apr 15 2004;350(16):1617-1628.

24. Breems DA, Boogaerts MA, Dekker AW, et al. Autologous bone marrow transplantation as consolidation therapy in the treatment of adult patients under 60 years with acute myeloid leukaemia in first complete remission: a prospective randomized Dutch-Belgian Haemato-Oncology Co-operative Group (HOVON) and Swiss Group for Clinical Cancer Research (SAKK) trial. *Br J Haematol.* Jan 2005;128(1):59-65.

25. Lowenberg B, Boogaerts MA, Daenen SM, et al. Value of different modalities of granulocyte-macrophage colony-stimulating factor applied during or after induction therapy of acute myeloid leukemia. *J Clin Oncol.* Dec 1997;15(12):3496-3506.

26. Lowenberg B, van Putten W, Theobald M, et al. Effect of priming with granulocyte colony-stimulating factor on the outcome of chemotherapy for acute myeloid leukemia. *N Engl J Med.* Aug 21 2003;349(8):743-752.

27. Ossenkoppele GJ, Graveland WJ, Sonneveld P, et al. The value of fludarabine in addition to ARA-C and G-CSF in the treatment of patients with high-risk myelodysplastic syndromes and AML in elderly patients. *Blood.* Apr 15 2004;103(8):2908-2913.

28. Schlenk RF, Benner A, Hartmann F, et al. Risk-adapted postremission therapy in acute myeloid leukemia: results of the German multicenter AML HD93 treatment trial. *Leukemia.* Aug 2003;17(8):1521-1528.

29. Schlenk R, Döhner K, Mack S, et al. Prospective evaluation of allogeneic hematopoietic stem cell transplantation from matched related and matched unrelated donors in younger adults with high-risk acute myeloid leukemia: Results of German-Austrian AMLSG treatment trial AMLHD98A. *J Clin Oncol in press.* 2010.

30. Heil G, Krauter J, Raghavachar A, et al. Risk-adapted induction and consolidation therapy in adults with de novo AML aged ≤ 60 years: results of a prospective multicenter trial. *Ann Hematol.* Jun 2004;83(6):336-344.

31. Care RS, Valk PJ, Goodeve AC, et al. Incidence and prognosis of c-KIT and FLT3 mutations in core-binding factor (CBF) acute myeloid leukaemias. *Br J Haematol.* Jun 2003;121(5):775-777.

32. Valk PJ, Bowen DT, Frew ME, Goodeve AC, Lowenberg B, Reilly JT. Second hit mutations in the RTK/RAS signaling pathway in acute myeloid leukemia with inv(16). *Haematologica.* Jan 2004;89(1):106.

33. Verhaak RG, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood.* Dec 1 2005;106(12):3747-3754.

34. Kohlmann A, Bullinger L, Thiede C, et al. Gene expression profiling in AML with normal karyotype can predict mutations for molecular markers and allows novel insights into perturbed biological pathways. *Leukemia.* Jun;24(6):1216-1220.

35. Cheson BD, Bennett JM, Kopecky KJ, et al. Revised recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia. *J Clin Oncol.* Dec 15 2003;21(24):4642-4649.

36. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976;63:655-660.

37. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics.* 1979;6(2):65–70.

38. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biom J.* Feb;52(1):70-84.

39. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* 1996;58(1):267-288.

40. Pabst T, Eyholzer M, Haefliger S, Schardt J, Mueller BU. Somatic CEBPA mutations are a frequent second event in families with germline CEBPA mutations and familial acute myeloid leukemia. *J Clin Oncol.* Nov 1 2008;26(31):5088-5093.

41. Renneville A, Mialou V, Philippe N, et al. Another pedigree with familial acute myeloid leukemia and germline CEBPA mutation. *Leukemia.* Apr 2009;23(4):804-806.

CHAPTER 4

SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels

Mathijs A. Sanders[1], Roel G.W. Verhaak[1,2], Wendy M.C. Geertsma-Kleinekoort[1], Saman Abbas[1], Sebastiaan Horsman[3], Peter J. van der Spek[3], Bob Löwenberg[1], and Peter J.M. Valk[1]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands

[2] Dana-Farber Cancer Institute, Department of Medical Oncology, Harvard Medical School, Boston, USA; The Broad Institute of M.I.T. and Harvard, Cambridge, USA

[3] Erasmus University Medical Center, Department of Bioinformatics, Rotterdam, The Netherlands

BMC Genomics 2008 Jan 25;9:41

ABSTRACT

Background. Accurate analyses of comprehensive genome-wide single nucleotide polymorphism (SNP) genotyping and gene expression data sets is challenging for many researchers. In fact, obtaining an integrated view of both large scale SNP genotyping and gene expression is currently complicated since only a limited number of appropriate software tools are available.

Results. We present SNPExpress, a software tool to accurately analyze SNP genotype calls, copy numbers, polymorphic copy number variations and gene expression in a combinatorial and efficient way. In addition, SNPExpress allows concurrent interpretation of these items with Hidden Markov Model inferred loss-of-heterozygosity and copy number regions.

Conclusion. The combined analyses with the easily accessible software tool SNPExpress will not only facilitate the recognition of recurrent genetic lesions, but also the identification of critical pathogenic genes.

INTRODUCTION

High-density genome-wide views of biological samples using high-throughput DNA mapping and mRNA gene expression microarrays facilitate the identification of recurrent molecular lesions. Both types of microarrays, which are being produced by different manufacturers, e.g., Nimblegen, Agilent, Sequenom, Applied Biosystems, Illumina and Affymetrix, typically contain large numbers of small oligonucleotides that interrogate the genome. Currently available DNA arrays contain over 500.000 probe sets, while the gene expression arrays target over 20.000 genes. Efficient analysis of these large datasets remains a challenge for many researchers.

The Affymetrix and Illumina DNA mapping platforms have been designed to specifically target sequences containing single nucleotide polymorphisms (SNPs). SNPs are currently estimated to be present at a frequency of 1 out of 300 nucleotides.[1] By including different probe sets to detect the possible SNP variants, genome-wide genotyping is feasible and these platforms can easily be applied to determine copy numbers of these chromosomal markers, similar to array comparative genomic hybridization (CGH). Due to the high number of SNPs, sample DNA can be examined with an inter-marker distance of 6 to 12 kb, and (micro) deletions and/or amplifications are detectable. By comparing disease samples to normal germ line DNA, a detailed overview of acquired gains and losses of the genome is obtained. In fact, although our knowledge is still developing, it has recently become apparent that that copy number variation (CNV) accounts for a substantial amount of genetic variation in the human genome.[2]

The high-resolution scanning technologies enable the analyses of CNV and associated phenotypes.[2] The power of DNA mapping has been shown extensively in cancer research. Chromosomal gains and losses as well as regions of loss-of-heterozygosity (LOH) have been shown in, for instance, leukemia[3,4], lung cancer[5-7] and colon cancer.[8] Recognition of recurrent lesions will ultimately result in the identification of pathogenic genes. For instance, SNP array analysis of a set of cancer cell lines has led to the identification of the microphthalmia-associated transcription factor *MITF* as a melanoma oncogene.[9]

On the Illumina platform genotypes are determined using hybridization of genomic DNA to BeadChips followed by an enzymatic discrimination step. On the Affymetrix platform, genotype calls and copy numbers are determined by a probe set. In analogy with the expression probe set, the genotype and copy number of an individual SNP is dependent on the balance of genotype calls in the associated probe set. Several methods for genotype calling[10-13] and assessment of copy number[14,15] have been developed. Advanced analysis methods of DNA mapping array data have focused on the identification of regions of LOH, or gains and losses.[16-19]

A particular SNP genotype or numerical changes in chromosome copy number can have profound effects on gene expression. A possible relation to tumor development was shown in breast cancer, where a 17q23 amplification was related to increased expression of genes at this locus[20] and in acute myeloid leukemia (AML), where amplification of 8p24 was associated with

increased expression of genes such as *MYC*.[21] In fact, SNPs as well as CNVs have recently been shown to have consistent effects, often in *cis*, on gene expression.[22,23] The integrated analysis of gene expression and SNP array data is a prerequisite to recognize these effects. To our knowledge, only one software package is able to visualize chromosome copy number and gene expression levels.[17] Here, we present a package, SNPExpress, which allows concurrent interpretation of genotype, HMM inferred LOH regions, CNVs, HMM inferred copy number and gene expression data. Due to the simple format of the input data, our package is not restricted to specific methods to determine genotype, copy numbers or expression level. Little knowledge of software is necessary to use SNPExpress, making the tool widely accessible.

IMPLEMENTATION

SNPExpress, written in JAVA (version 1.5), uses tab-delimited files as input and is currently available for use with Affymetrix DNA mapping arrays (10K 2.0, 100K and 500K set), Illumina HumanHap550 Genotyping BeadChip and Affymetrix GeneChips (HG-U95Av2, HG-U133A and B, HG-U133 plus 2.0). A file containing a matrix with each column representing the genotypes of one array and rows starting with Illumina or Affymetrix SNP IDs is mandatory. The genotype should be formatted as homozygous 'AA' or 'BB', heterozygous 'AB', or, 'noCall' (Affymetrix)/ 'NC' (Illumina). Similar matrix files containing copy numbers or gene expression values are optional. Copy numbers should be centered around 2, where 2 represents the normal copy number of the autosomes and 1 for the male X chromosome. The maximum displayed copy number is 4, in case the copy number is above 4 this is indicated by a darkblue background. Copy number-, genotype- and gene expression files required for SNPExpress can be generated through tools such as Affymetrix BRLMM[13], GCOS/CNAT 4.0[24], or dChipSNP.[17] In case of Illumina data, SNPExpress includes the non-synonymous SNPs and the MHC region, however, mitochondrial SNPs and Y-chromosome SNPs are not visualized. All files can be optionally uploaded as tab- or comma-delimited .txt files or binary files. These binary files can be created from .txt files by the menu item 'convert data source'.

SNPExpress maps both the SNP IDs (Illumina and Affymetrix) and the expression probe set IDs (Affymetrix) to the genome through internal alignment tables, using annotation provided by the manufacturer. Annotation was generated using NCBI build 36.1. Regions showing LOH are calculated through a Hidden Markov Model (HMM), described previously.[18] The probability values for heterogeneous calls required for the HMM have been generated through sets of genotypes of normal samples. For the 100K and 500K array, 90 samples and 270 samples, respectively, of different ethnical background from the HapMap project are available through the NCBI GEO website (and provided by the manufacturer).[28,29] For the 10K array normal matched blood samples available through the GEO public repository have been processed.[30] Since reference normal Illumina genotype datasets are currently not publicly available, LOH regions using this platform are not supported in this version of SNPExpress. SNPExpress includes the option to visualize the

results of a novel analytical method that infers the copy number of each SNP based on a HMM model, which is implemented in dChipSNP.[17,31] Also, all CNVs[2], currently cataloged in the Database of Genome Variants[32], can be visualized.

Example expression, copy number, genotype and HMM copy number example files of two AML patients can be downloaded.[33]

RESULTS

Genotypes and copy numbers are displayed as sequential blocks by which color indicates genotype, horizontal coordinate indicates position on the chromosome and vertical coordinate indicates copy number (Figure 1). The colored genotype blocks are drawn sequentially in a chromosome-wide view and proportional to chromosomal location when zoomed into a region of interest. Gene expression levels are visualized as a vertical bar at the chromosomal position of the gene-specific probe set. The height of the bar is proportional to the gene expression value. The default value is 500 and expression higher than 500 is capped at 500, however, these values are user-definable. In the event that multiple probe sets span the same region in the chromosome-wide view the vertical gene expression bars are red and proportional to the highest expression value. Zooming into the location of interest discloses the individual probe sets. Links of SNP IDs to public databases are available by holding the ctrl-key and clicking on a SNP ID.

Distinct background colors are used to accentuate genomic changes. Individual copy numbers are indicated as gain (pink background) or loss (green background) when their value exceeds a user-defined value. The default deviation threshold is 0.5. LOH is highlighted at diploid level by a bold magenta line (Figure 1). All colors can be adapted to the users' preferences.

From the menu, the user is able to choose to visualize either one chromosome of multiple samples or the complete genome of one sample. Detailed information, such as SNP ID, associated gene symbol, probe set ID, cytoband and expression value, is shown on a mouse-over display. Furthermore, a gene of interest is directly visualized through a search function, and its associated SNPs are indicated with an orange background color. The options to display known CNVs (purple background) or the HMM copy number results (thin magenta line) are included (Figure 1C). Finally, relevant data of a particular minimal deleted of amplified region can be exported (i.e., Sample, Probe_set_id, Chromosome, Location (bp), Cytoband, Associated gene, Genotype, Copy number and Inferred LOH of the selected region) and high-resolution images of the visualization can be saved in the Portable Network Graphic (PNG) format.

To illustrate the power of SNPExpress, DNA mapping array profiles of tumor samples of a series of 48 AML patients were generated using Affymetrix 250K *Nsp*I DNA mapping arrays. Ficoll separation

of the mononuclear cells from AML typically yields >80% pure population of leukemic blast cells. High molecular weight DNA was isolated from these malignant cells and the Affymetrix mapping arrays were used according to the protocol of the manufacturer. Genotypes were calculated using BRLMM and copy numbers were assessed using dChipSNP. Biotin-labeled cRNA of the same AML samples was hybridized on Affymetrix HG-U133 plus 2.0 GeneChips, as previously described.[34] The resulting dataset was imported in SNPExpress for analyses. Large chromosomal regions showing loss or gains of genetic material are known to be apparent in leukemic blasts of AML patients. Well-known examples of chromosomal lesions in AML are monosomies of chromosome 5 and 7, which have been associated with a poor prognosis.[35] Using SNPExpress, monosomies of chromosome 7 were evidently demonstrated in AML samples, previously shown by cytogenetics (Figure 1). SNPExpress also correctly predicted the presence of LOH as a result of the absence of one chromosome 7. In fact, 17 out of 21 numerical cytogenetic aberrations, i.e., whole chromosomes and interstitial deletions, in 48 AML samples analyzed, were recognized by using SNPExpress. Four numerical abnormalities, present in less than 30% of the AML cells, were missed. Chromosomal gains, losses as well as uniparental disomy (UPD) may also have other important consequences, such as affecting expression of (imprinted) genes. Combinatorial visualization of genotype, copy number and gene expression is a prerequisite to recognize these aberrations. For example, the majority of genes show located on chromosome 7 show an overall decrease in expression in AML cases with a monosomy 7 (Figure 1).

Large regions of homozygosity are present in approximately 20% of primary AML cases as a result of segmental UPD.[3,36] These regions of UPD seemed to be non-random and may be used to unmask pre-existing recessive mutations in leukemia genes, such as *CEBPA*, *WT1*, *FLT3* and *RUNX1*.[3,37] SNPExpress adequately identified regions of UPD involving e.g. chromosome 11p (Figure 1D), in two patients with a normal karyotype. UPD involving chromosome 11 is associated with homozygous mutations in *WT1*.[37] Interestingly, in 13 out of 48 AML patients (27%) large regions of segmental UPD continuing to the telomere were recognized using SNPExpress.

These examples demonstrate the power of SNPExpress. To our knowledge, no tool is currently available that allows concurrent interpretation of genotype, HMM inferred LOH regions, copy number, HMM inferred copy number and gene expression data. Moreover, no specialized knowledge is necessary to work with SNPExpress.

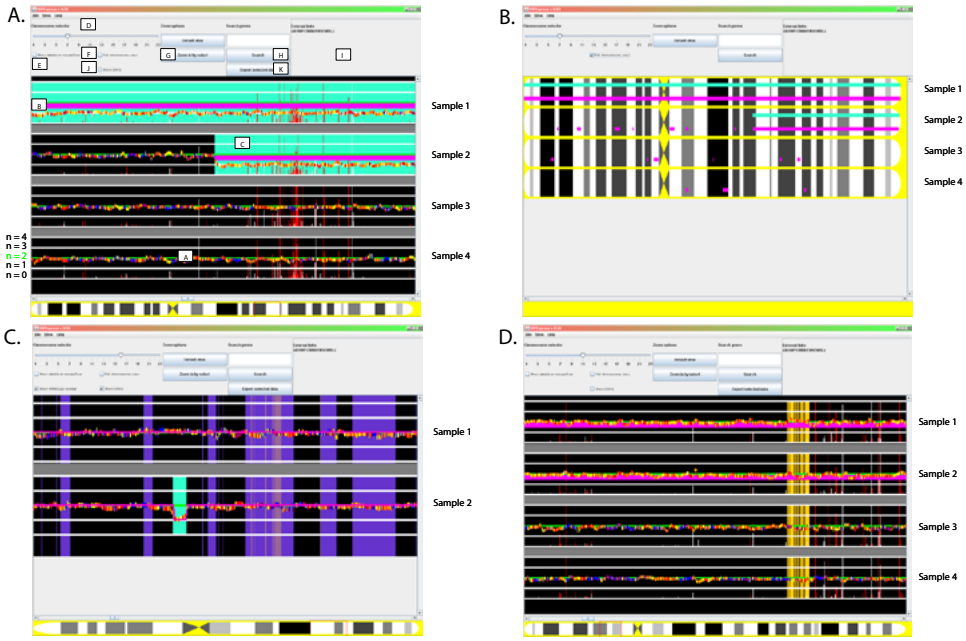


Figure 1. SNPExpress screenshot. A. DNA mapping array data from the Affymetrix 250K *Nsp*I DNA mapping array depicting chromosome 7 of four AML samples. Copy numbers (n=0, 1, 2, 3, 4) are shown for each individual patient by horizontal lines. (A) SNP genotypes are sequentially aligned along the chromosome (AA: red; BB: yellow; AB: blue, noCall: white). LOH is indicated by a thick magenta horizontal bar (B), gains (default n>2.5) by a pink and losses (default n<1.5) by a turquoise background (C). Gene expression levels are visualized as vertical white bars. Multiple probe sets spanning the same locus is depicted by a red vertical gene expression bar proportional to the highest expression value. The two upper samples display a complete monosomy (sample 1) or a deletion of the q-arm of chromosome 7 (sample 2). The chromosome selecter (D; where 23 is the X chromosome), the mouse-over function showing info of each SNP or probe set (E), full chromosome view (F), zoom function (G) gene search function (H), the links to external databases (I), display CNVs (J) and export selected data (K) options are indicated. **B.** Full chromosome view of samples from Figure 1A. **C.** CNV (purple background) and copy number of each SNP based on a HMM model (HMM copy number, magenta line) for two AML cases. **D.** An example of large scale UPD on chromosome 11 in the upper two AML patients in comparison to two other AML samples. The overall copy number is two and large regions of LOH are indicated by the thick magenta line across the chromosome. SNPs associated with *WT1* are depicted with an orange background.

DISCUSSION

Since genome-wide DNA mapping array and mRNA expression studies become more cost effective, the number of samples profiled on these platforms will increase. Specialized user-friendly tools for efficient visualization, such as SNPExpress, will therefore be indispensable. In fact, the initial version of SNPExpress has already been successfully applied in showing segmental uniparental disomy as a recurrent mechanism for homozygous *CEBPA* mutations in acute myeloid leukemia.[38]

Other tools for visualizing and processing SNP array data, such as SNPScan[39], SIGMA[40], ArrayFusion[41], Partek Genomics Suite[42] and GenePattern[43] have been developed. Most of these tools incorporate visualization options for displaying LOH (GenePattern, Partek Genomics Suite, SNPScan) and copy number (all but ArrayFusion), whereas SNPScan and ArrayFusion have output functionality that facilitates linking SNP data to the UCSC genome browser.[39,41] Some are linked to a private database, which restricts pre-processing of the array data, but gives the advantage of data storage.[40] GenePattern and the Partek Genomics Suite provide normalization and data smoothing functionality. These two packages and SNPScan have also incorporated options for combined analysis of paired samples, i.e., tumor and normal. Like SNPExpress, SNPscan, GenePattern, and the Partek Genomics Suite can detect regions of LOH, amplification and deletion. None of these tools describe the ability to process Illumina BeadArray files. Where SNPExpress may lack the opportunity to directly process raw data files (such as Affymetrix CEL-files), it adds integrated visualization of expression (Affymetrix) and DNA copy number and genotype (Affymetrix and Illumina) data. Moreover, we believe that this is provided in a user-friendly way that does not require specialist computer knowledge.

SNPExpress has some limitations. A full-length chromosome view depicting gains, losses and the regions showing LOH is feasible using SNPExpress. However, the large datasets generated by the 500K mapping array platform makes it impossible to visualize the sequentially aligned SNPs of the full-length chromosomes on one screen. Selecting the most informative SNPs, i.e., representative for particular haplotypes, may solve this issue. Such algorithms are currently in development. Furthermore, the current implementation of the HMM could also be improved by implementing a HMM that takes into account the effects of linkage disequilibrium, i.e., LD-HMM.[18] The number of samples to be visualized concurrently is limited by the memory available to the application.

CONCLUSIONS

The power of SNPExpress, as with previously developed tools[44], is its high accessibility and powerful visualization, which facilitates the identification of biologically and clinically relevant entities. We have shown that recurrent biologically relevant entities, such as chromosomal gains or losses and LOH in AML, are accurately identified with SNPExpress. Hence, SNPExpress will be beneficial to genome-wide studies by providing an integrated view of data from DNA mapping and mRNA expression arrays in an easily accessible and accurate way.

Authorship contributions

MAS wrote and designed the software; RGWV designed the software, performed the analysis and wrote the manuscript; WGK performed experiments; SA gave intellectual contributions; SH contributed code; PJS gave intellectual contributions; BL gave intellectual contributions; PJMV

designed the study, gave intellectual contributions and wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgments

The research described was supported by grants from the Erasmus University Medical Center (Revolving Fund) and the Dutch Cancer Society "Koningin Wilhelmina Fonds". We are indebted to Andy Hall for providing Affymetrix 10K DNA mapping array data at the initial set up of SNPExpress.

REFERENCES

1. International HapMap Consortium: A haplotype map of the human genome. *Nature*. 2005;437(7063):1299-1320.
2. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: new insights in genome diversity. *Genome Res*. 2006;16(8):949-961.
3. Raghavan M, Lillington DM, Skoulakis S, et al. Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Res*. 2005;65(2):375-378.
4. Irving JA, Bloodworth L, Bown NP, Case MC, Hogarth LA, Hall AG. Loss of heterozygosity in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis. *Cancer Res*. 2005;65(8):3053-3058.
5. Zhao X, Weir BA, LaFramboise T, et al. Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res*. 2005;65(13):5561-5570.
6. Lindblad-Toh K, Tanenbaum DM, Daly MJ, et al. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat Biotechnol*. 2000;18(9):1001-1005.
7. Janne PA, Li C, Zhao X, et al. High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. *Oncogene*. 2004;23(15):2716-2726.
8. Nakao K, Mehta KR, Fridlyand J, et al. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*. 2004;25(8):1345-1357.
9. Garraway LA, Widlund HR, Rubin MA, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*. 2005;436(7047):117-122.
10. Di X, Matsuzaki H, Webster TA, et al. Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics*. 2005;21(9):1958-1963.
11. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics*. 2006;22(1):7-12.
12. Lamy P, Andersen CL, Wikman FP, Wiuf C. Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res*. 2006;34(14):e100.
13. Affymetrix: BRLMM: an Improved Genotype Calling Method for the GeneChip® Human Mapping 500K Array Set. *In Santa Clara, CA*. 2006:1-18. (http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf).
14. Nannya Y, Sanada M, Nakazaki K, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res*. 2005;65(14):6071-6079.
15. Huang J, Wei W, Zhang J, et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum Genomics*. 2004;1(4):287-299.
16. LaFramboise T, Weir BA, Zhao X, et al. Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Comput Biol*. 2005;1(6):e65.
17. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*. 2004;20(8):1233-1240.
18. Beroukhim R, Lin M, Park Y, et al. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput Biol*. 2006;2(5):e41.
19. Huang CC, Taylor JM, Beer DG, Kardia SL. Hidden Markov model for defining genomic changes in lung cancer using gene expression data. *OMICS*. 2006;10(3):276-288.
20. Monni O, Barlund M, Mousses S, et al. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. *Proc Natl Acad Sci U S A*. 2001;98(10):5711-5716.
21. Rucker FG, Bullinger L, Schwaenen C, et al. Disclosure of candidate genes in acute myeloid leukemia with complex karyotypes using microarray-based molecular characterization. *J Clin Oncol*. 2006;24(24):3887-3894.

22. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet*. 2007;39(10):1217-1224.

23. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315(5813):848-853.

24. Affymetrix (2007). GCOS. Retrieved 1 June 2007, from http://www.affymetrix.com/support/technical/product_updates/gcos_download.affx.

25. Illumina (2007). Annotation HumanHap550 Genotyping BeadChip. Retrieved 1 June 2007, from http://www.illumina.com/pages.ilmn?ID=154.

26. Affymetrix (2007). Mapping Array Annotation. Retrieved 1 June 2007, from http://www.affymetrix.com/support/technical/byproduct.affx?cat=dnaarrays.

27. Affymetrix (2007). Expression array probe set alignments. Retrieved 1 June 2007, from http://www.affymetrix.com/Auth/analysis/downloads/psl/HG-U133_Plus_2.link.psl.zip.

28. Affymetrix (2007). Reference dataset Affymetrix 100K Mapping Array. Retrieved 1 June 2007, from http://www.affymetrix.com/support/technical/sample_data/hapmap_trio_data.affx.

29. Affymetrix (2007). Reference dataset Affymetrix 500K Mapping Array. Retrieved 1 June 2007, from http://www.affymetrix.com/support/technical/sample_data/500k_data.affx.

30. Affymetrix (2007). Reference dataset Affymetrix 10K Mapping Array. Retrieved 1 June 2007, from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2959.

31. Zhao X, Li C, Paez JG, et al. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*. 2004;64(9):3060-3071. 32. Genomics TCfA (2007). Database of Genome Variants. Retrieved 1 June 2007, from http://projects.tcag.ca/variation.

33. Sanders MA (2007). Homepage SNPExpress. Retrieved 1 June 2007, from http://www.planetmathematics.com/SNPExpress/.

34. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1617-1628.

35. Mrozek K, Heerema NA, Bloomfield CD. Cytogenetics in acute leukemia. *Blood Rev*. 2004;18(2):115-136.

36. Gorletta TA, Gasparini P, D'Elios MM, Trubia M, Pelicci PG, Di Fiore PP. Frequent loss of heterozygosity without loss of genetic material in acute myeloid leukemia with a normal karyotype. *Genes Chromosomes Cancer*. 2005;44(3):334-337.

37. Fitzgibbon J, Smith LL, Raghavan M, et al. Association between acquired uniparental disomy and homozygous gene mutation in acute myeloid leukemias. *Cancer Res*. 2005;65(20):9152-9154.

38. Wouters BJ, Sanders MA, Lugthart S, et al. Segmental uniparental disomy as a recurrent mechanism for homozygous CEBPA mutations in acute myeloid leukemia. *Leukemia*. 2007;21(11):2382-2384.

39. Ting JC, Ye Y, Thomas GH, Ruczinski I, Pevsner J. Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. *BMC Bioinformatics*. 2006;7:25.

40. Chari R, Lockwood WW, Coe BP, et al. SIGMA: a system for integrative genomic microarray analysis of cancer genomes. *BMC Genomics*. 2006;7:324.

41. Yang TP, Chang TY, Lin CH, Hsu MT, Wang HW. ArrayFusion: a web application for multi-dimensional analysis of CGH, SNP and microarray data. *Bioinformatics*. 2006;22(21):2697-2698.

42. Partek (2007). Partek Discovery Suite. Retrieved 1 June 2007, from http://www.partek.com.

43. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nat Genet*. 2006;38(5):500-501.

44. Verhaak RG, Sanders MA, Bijl MA, et al. HeatMapper: powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics. *BMC Bioinformatics*. 2006;7:337.

CHAPTER 5

Detailed genome analyses reveal extensive RAG-mediated rearrangements and aberrations in *NF1* and *SUZ12* in adult acute leukemia subsets

Mathijs A. Sanders[1], Remco Hoogenboezem[1], Carla Exalto[1], Annikó Szabo[1], Annelieke Zeilemaker[1], Marta Pratcorona[2], Jasper E. Koenders[1], Anita Schelen[1], Peter van Geel[1], H. Berna Beverloo[3], Jan Cornelissen[1], Anita W. Rijneveld[1] and Peter J.M. Valk[1]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands
[2] IDIBAPS, Hospital Clínic de Barcelona, Department of Hematology, Barcelona, Spain
[3] Erasmus University Medical Center, Department of Clinical Genetics, Rotterdam, the Netherlands

Work in progress

ABSTRACT

Acute leukemia is characterized by the accumulation of somatic genetic alterations in hematopoietic stem and progenitor cells eventually resulting in overt leukemia. To characterize somatic events associated with leukemic transformation, we performed detailed genome-wide copy number analyses of 53 adult B-cell acute lymphoblastic leukemia (ALL), 20 adult T-cell ALL (T-ALL) and 100 adult acute myeloid leukemia (AML) cases. We observed recurrent aberrations involving *CDKN2A/B*, *IKZF1*, *PAX5* and *BTG1* at relatively high frequencies, with the *CDKN2A/B* pathway being perturbed in all adult T-ALL cases. In adult AML, focal copy number alterations were virtually lacking. Interestingly, genetic lesions simultaneously perturbing the genes encoding for *NF1*, involved in RAS pathway inhibition, and *SUZ12*, a pivotal polycomb repressive complex 2 (PRC2) member, were recurrently found in T-ALL and AML. Gene expression profiling (GEP) analysis revealed the substantial down regulation of both genes supporting the notion that PRC2 loss cooperates with RAS pathway activation in acute leukemia. Finally, targeted resequencing of regions harboring recurrent genetic alterations in 5 selected B-ALL cases revealed extensive involvement of the recombination activating genes (RAG) complex as a mutational mechanism invoking large deletions as well as complex insertions and deletions in promoters, enhancers and open chromatin proximal to genes regulating B-cell development.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter5/

INTRODUCTION

Acute lymphoblastic leukemia (ALL) is a heterogeneous disease characterized by a clonal proliferation of malignant lymphoblasts, initiated by the accumulation of genetic alterations in B- or T-lymphoid precursor cells. ALL is the most common leukemia in children (26% of all cancer types and 80% of all paediatric leukemias), but is a rare disease in adults (2-3% of all cancer types and 56% of all adult acute leukemias).[1-3] With the introduction of different genome characterization methodologies much has been learned about the molecular biology of ALL leading to risk-adapted treatments, however, the long-term survival of adults is still inferior to that of children (5-year event-free survival rate 80% versus 40%, respectively).[1-3] Cytogenetic characterization of adult ALL revealed recurrent interchromosomal aberrations involving *BCR-ABL1, E2A-PBX1* and *MLL*-rearrangements in B-cell ALL (B-ALL), and *SIL-TAL1* in T-cell ALL (T-ALL).[4] Genetic analysis of especially paediatric ALL has been extensively carried out, while the genome of adult ALL remains scarcely characterized. Array-based genome characterization of predominantly paediatric ALL has revealed frequent and recurrent genetic alterations, mostly comprising deletions of promoters or complete genes involved in the development of B- or T-lymphoid precursor cells[5,6], such as deletions affecting *CDKN2A/B*, *IKZF1*, *PAX5*, *EBF1* and *BTG1*. The adaptive immune system requires diversification for the defence against the invasion of variegated pathogens. This defence is conferred by the diversification of antibody production by cutting and recombining variable (V), diversification (D) and joining (J) gene segments by a process named V(D)J recombination. Previous studies have postulated that many of the recurrent deletions may be effectuated by illegitimate V(D)J recombination utilizing cryptic recombination signal sequence (RSS) motifs in the vicinity of the nascent copy number alteration (CNA) breakpoint boundaries.[7,8] Recent efforts have shown that a substantial number of CNAs are flanked by cryptic RSS motifs in specifically paediatric *ETV6-RUNX1* ALL patients[9], especially in regions marked by active histone markers, such as H3K4me3. The RAG complex mediates V(D)J recombination within recombination centre foci located in antigen receptor loci, characterized by an enrichment of H3K4me3, H3 acetylation and RNA polymerase II binding.[10] Altogether, these studies culminate into the hypothesis postulating that the RAG endonuclease complex is aberrantly targeted to loci accommodating cryptic RSS motifs and subsequently invoking illegitimate genetic lesions perturbing developmental genes associated with leukemic and clonal evolution.

Acute myeloid leukemia (AML) is a malignant and heterogeneous disease characterized by the acquisition of genetic lesions in hematopoietic stem and progenitor cells.[11] Cytogenetic characterization has revealed recurrent interchromosomal aberrations, e.g. t(15;17), t(8;21), inv(16) and *MLL*-rearrangements, which have consecutively been used for prognostication. Recent studies adopting next generation sequencing (NGS) technologies have revealed recurrent molecular aberrations involving genes associated with transcription activity[12], epigenetic modifications[13,14], spliceosome machinery[15], and cohesion complex formation.[16] The prognostic

significance of many of these newly identified molecular aberrations remains equivocal. Array-based characterization of the AML genome revealed the general scarcity of CNAs precluding the identification of many novel genes involved in leukemogenesis[17] and ostensibly highlighting its stability.

In order to determine the frequency of recurrent aberrations and to identify novel genetic aberrations in adult acute leukemia, we have characterized 53 B-ALL, 20 T-ALL, and 100 AML cases at diagnosis by copy number variation (CNV) analysis. Additionally, we determined which genetic lesions are acquired, potentially driving leukemic transformation in concert.

Subsequently, the CNV analyses revealed that a substantial fraction of the adult ALL cases acquired multiple focal deletions located proximal to the transcription start site (TSS) or involving the complete promoter of genes associated with development and differentiation processes in B- or T-lymphoid precursor cells.[5,6] We selected 5 B-ALL cases substantially exhibiting this behaviour and with the advent of NGS demonstrated that almost all deletion events are flanked by cryptic RSS motifs.

METHODS

Patients samples

After informed consent, bone marrow aspirates or peripheral blood samples of a representative cohort of adult ALL and AML patients were collected. Eligible patients had a diagnosis of primary ALL or AML, confirmed by cytological examination and immunophenotyping of blood and bone marrow. The majority of these cases were treated following the HOVON (Dutch-Belgian Hematology-Oncology Co-operative group) protocols (http://www.hovon.nl). Blasts and mononuclear cells were purified by Ficoll-Hypaque (Nygaard, Oslo, Norway) centrifugation and cryopreserved. All samples contained 80-to-100 percent blast cells after thawing, regardless of the blast count at diagnosis.

Gene expression profiling

RNA was isolated from 136 B-ALL, 55 T-ALL and 661 AML adult cases using RNABee. Gene expression profiles of the samples were generated using Affymetrix HG-U133 plus 2.0, as described elsewhere.[18] Gene expression data is available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress), see Supplementary material. The different cohorts have been made comparable by mean-centering the expression levels to 0 per dataset. Differential gene expression analysis was performed with the Mann–Whitney U test in the R environment. All additional plots were generated by ggplot2.[19]

Array based copy number analysis

Genome-wide genotyping data of 73 ALL patients, i.e., 53 B-ALL and 20 T-ALL diagnostic samples in conjunction with paired remission samples, were generated using Affymetrix 6.0 *Nsp*I/*Sty*I DNA mapping arrays and for the 100 AML diagnostic samples alone with the Affymetrix 500K *Nsp*I/*Sty*I DNA mapping arrays. DNA mapping array data is available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress), see Supplementary material. High-molecular-weight DNA was isolated using columns (Qiagen, Hilden, Germany) and the Affymetrix DNA mapping arrays were used according to the protocol of the manufacturer (Affymetrix, Santa Clara, CA). In brief, 250 ng of genomic DNA was digested with *Nsp*I or *Sty*I and ligated to *Nsp*I or *Sty*I adapters using T4 DNA ligase (New England Biolabs, Ipswich, MA, USA). Samples were then amplified by PCR using TITANIUM Taq polymerase (Clontech, Mountain View, CA, USA). PCR products were pooled and purified using the Clontech purification kit and subjected to fragmentation using DNaseI. The DNA fragments were subsequently biotin-labeled with terminal deoxynucleotidyl transferase, hybridized on the array in a GeneChip® Hybridization Oven 640, washed and stained in a GeneChip® Fluidics Station 450. Data was obtained using the GeneChip Scanner 3000 7G.

Genotypes were calculated using the Birdseed algorithm and copy numbers were assessed using dChip.[20] The copy numbers of all AML samples were calculated using diploid references. CNV profiles of the ALL and AML cases where manually scrutinized with SNPExpress[21] and aberrations were selected when displaying constitutive loss or gain, i.e., 10 probe sets or more, and not observable in matched remission samples or not observable in all leukemia cases, e.g., copy number polymorphisms or systematic bias in copy number. Subclonal CNVs were selected only when 20 or more probe sets displayed constitutive reduced or increased copy number levels. Genes afflicted by CNAs were ranked according to frequency, but due to hyperdiploidy and chromosomal losses many passenger genes were simultaneously identified. To ascertain genes perturbed by or directly flanking CNAs we have utilized kernel density estimation with a flat-top Gaussian kernel distribution to account for the unequidistant nature of the probe sets.[22] In brief, the region encompassing the CNA is weighted equally, while flanking regions are weighted according to an exponential decay function based on distance and size of the CNA. To prevent the substantial weighting of very small aberration we have put a gamma distribution as a conjugate prior on the kernel distribution.

PCR, nucleotide sequencing and denaturing high performance liquid chromatography

All PCR reactions were carried out in the presence of 25mM dNTP, 15 pmol primers, 2mM $MgCl_2$, Taq polymerase and 1xbuffer (Invitrogen Life Technologies, Breda, The Netherlands) at an annealing temperature of 60^{0}C. The 16 exons of the *SUZ12* were amplified using the primers indicated in Supplementary Table 1 and the promoter of *BTLA* by a forward primer

(5'-GAGCCTGGATGATTTTGTGAA-3') and a reverse primer (5'-CCGTGACATGTACAGGAAAA-3'). Cycling conditions were: 1 cycle 5 min at 94^{0}C, 35 cycles 1 min at 94^{0}C, 1 min at 60^{0}C, 1 min at 72^{0}C, and 1 cycle 7 min at 72^{0}C. PCR products were purified using the Multiscreen-PCR 96-well system (Millipore, Bedford, MA) followed by direct sequencing with the appropriate forward and reverse primers using an ABI-PRISM3100 genetic analyzer (Applied Biosystems, Foster City, CA). All *SUZ12* PCR products of exon 14-15-16 or the *BTLA* promoter were subjected to denaturing high performance liquid chromatography (dHPLC) analyses using a Transgenomics (Omaha, NE) WAVE system. Samples were run at 55.6^{0}C.

Roche 454 next-generation sequencing

Amplicon sequencing was performed using the Roche GS Junior 454 system (Roche, Basel, Switzerland) following the protocols of the manufacturer. Sequence reads were processed and analyzed using the GS Amplicon Variant Analyzer (Roche, Basel, Switzerland). The *SUZ12* zinc finger and VEFS domain are encoded by exons 12-to-16. Primers are indicated in Supplementary Table 2). Amplicons carrying MIDs were generated and purified according to the Amplicon Library Preparation Method Manual (Roche, Basel, Switzerland). DNA enriched beads, carrying the amplification products, were generated according to the emPCR Amplification Method Manual-Lib-A (Roche, Basel, Switzerland); a beads-to-amplicon ratio of 1:2 was used.

Exome sequencing and targeted resequencing

From the diagnostic and remission material of 5 ALL cases the genomic DNA was sheared with the Covaris S2 (Covaris) with default settings for exome sequencing. Subsequently, the sample libraries were prepared using the TruSeq DNA Sample Preparation Guide (Illumina). The target chromosomal regions were derived from the DNA mapping array identified CNAs of the 5 ALL cases. The exons and the targeted regions (Supplementary Table 13) were captured by employing custom in-solution oligonucleotide baits (Nimblegen SeqCap EZ plus). The sample libraries were subjected to paired-end sequencing (2x100bp) on the HiSeq 2000 (Illumina) and were aligned against hg19 using the Burrows Wheeler Aligner (BWA)[23] with default settings. Reads aligning into undetermined regions of the human genome (hg19), e.g., segments of intron 3 belonging to the *IKZF1* gene, were aligned against the updated human genome sequence (hg38) containing these regions. Whole exome sequencing (WES) and targeted resequencing data derived from patient specimens have been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), see Supplementary material.

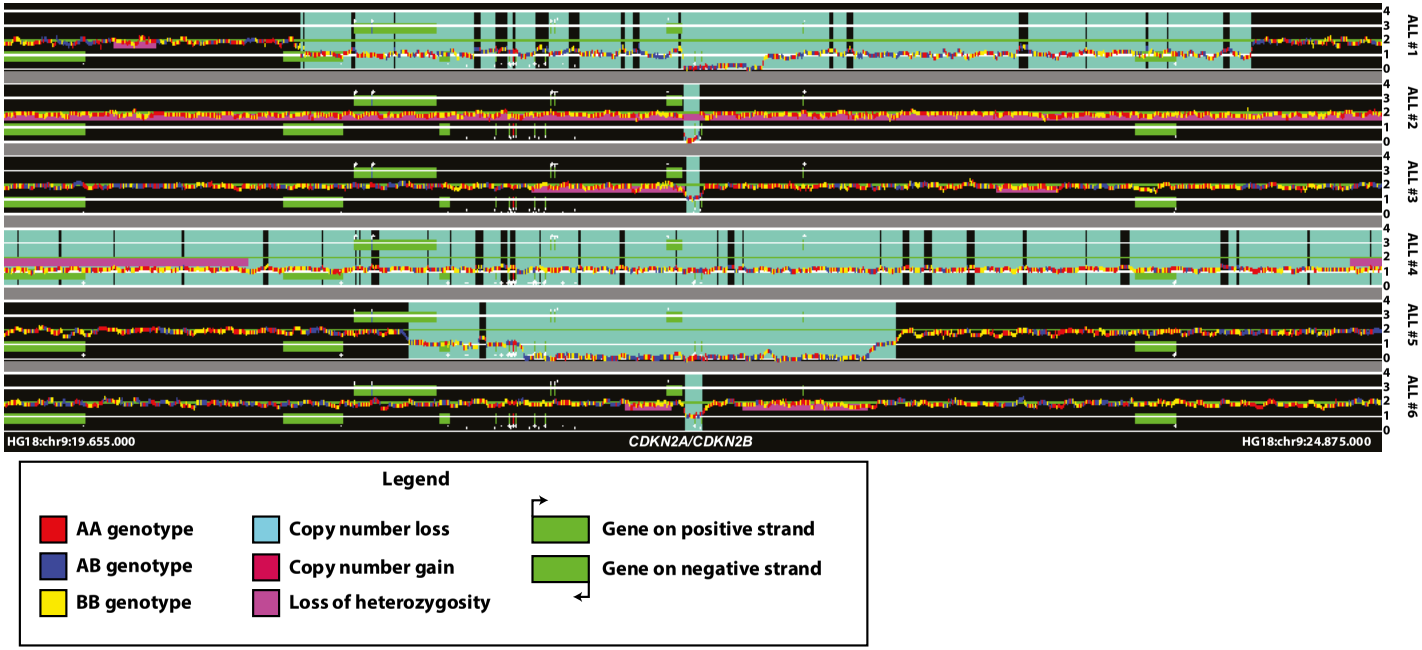


Figure 1. Copy number variation profiles of acute lymphoblastic leukemia patients. The copy number variation profiles of six ALL patients, specifically focusing on the region encompassing *CDKN2A and CDKN2B*. All six patients have acquired a heterozygous or homozygous loss of *CDKN2A* and *CDKN2B* with a high diversity of breakpoints.

Variant detection

Variants were determined with SAMtools[24], annotated with ANNOVAR[25], and filtered by an in-house developed algorithm. In brief, the algorithm compares the variant frequency of the diagnostic sample to a control sample, e.g., healthy tissue or remission material, taking into account the alignment quality of the reads and the local sequence context of the detected variant. Somatic complex mutations comprising a combination of insertions, deletions and mutations are detected with Pindel.[26]

Exact breakpoint and RSS detection

Copy number variation profiles from the NGS data of the 5 B-ALL cases were produced with CNVsvd (M.A.S., R.H, and P.J.M.V., manuscript in preparation). Exact breakpoint locations were determined with BreakDancer[27] (v1.12) or were manually scrutinized when undetected. Adjacent to either side of the breakpoint 150 base pair of sequence was extracted from hg19 using the SAMtools API.[24] To ascertain cryptic RSSs we uploaded all the breakpoint sequences to RSSsite[28] for detection of 12-bp spacer RSS or 23-bp spacer RSS based on the human detection algorithm. *De novo* motifs near the breakpoints were inferred by uploading all the breakpoints sequences to the MEME website.[29] The sequence logo for the detected cryptic RSS motifs by RSSsite were constructed with seqLogo.[30]

Histone markers and protein binding

We procured the histone marker data H3K4me3, H3K27ac, H3K27me3, RNA polymerase II for the B-lymphoblastoid cell line GM12878 from the Encyclopedia of DNA Elements (ENCODE) project.[31] Genome segmentation into 15 definable chromatin states, e.g. active promoter, based on combinations of epigenetic markers was constructed by the ENCODE project and procured from the UCSC website. Background probability of each chromatin state was determined by calculating the ratio of the summarized total length of all regions affiliated with this chromatin state with respect to the full genome length. To infer if RAG2 could bind in the vicinity of the detected CNA breakpoints we performed a UCSC liftOver from hg19 to mm9 to compare the breakpoint loci to Chip-Seq data of Rag2 binding in wild type murine thymocytes[10] (GEO omnibus GSM530318). Enrichment plots were generated with ngs.plot.[32]

RESULTS

Copy number variation analysis of adult ALL and AML

We have performed CNV analysis of diagnostic material of 53 B-ALL, 20 T-ALL and 100 AML cases (Table 1). In total we observed 1005 genomic alterations in ALL amounting to a mean of 13.77 genetic alterations per ALL case. In total we detected 268 genetic lesions in AML, amounting to

a mean of 2.68 genetic alterations per AML case. Subcategorization of recurrent genetic lesions demonstrate that certain aberrations are highly specific for AML or an ALL subtype (Table 2, Supplementary Table 3, and Supplementary Table 4). Common amongst B-ALL and T-ALL is the deletion of *CDKN2A* and *CDKN2B*[33] (Figure 1), 57% and 90% respectively, and *RB1,* 15% in both ALL subtypes. Strikingly, the two T-ALL cases without a deletion of *CDKN2A/B* have a deletion involving *CDKN2AIP* (*CDKN2A interacting protein*) and *CDKN2AIPNL* (*CDKN2A interacting protein N-terminal like*), implicating the perturbation of *CDKN2A/B* through different pathways. We detected genetic lesions highly specific for B-ALL involving *IKZF1* (47%), *PAX5* (36%), *CRLF2* (22%), *BTG1* (19%), *BTLA* (13%), *MKKS* (13%) and *EBF1* (6%). Additionally, we detected genetic lesions highly specific for T-ALL involving *NF1/SUZ12* (15%), *WT1* (15%) and unbalanced translocations affecting *TAL-1* (25%).

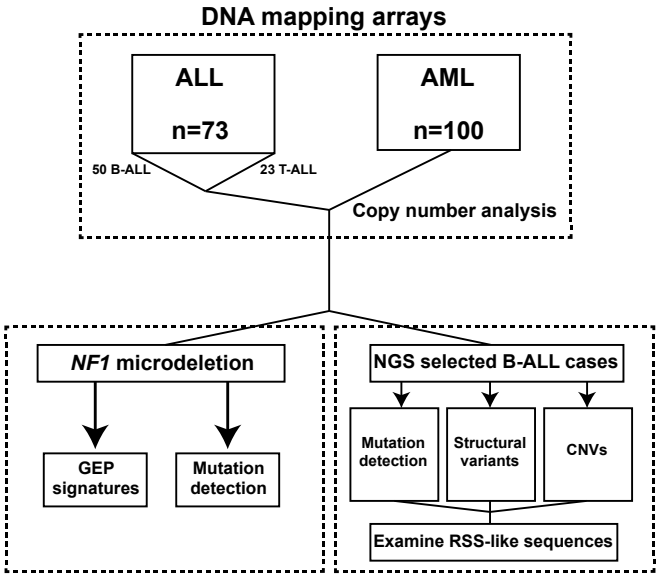


Figure 2. Analysis workflow for this study. Genomic characterization of acute lymphoblastic leukemia (53 B-ALL and 20 T-ALL) and acute myeloid leukemia by DNA mapping arrays is performed on each data set separately. AML and ALL cases with an acquired *NF1* microdeletion were compared to all other leukemia cases to determine if this subgroup is characterized by a GEP signature. A subset of B-ALL patients was selected, due to an increased number of focal deletions, for further genomic characterization. Mutations, structural variants and CNVs are detected for these patients and subsequently used for the determination of RSS-like sequences near breakpoints. Next generation sequencing (NGS), gene expression profiling (GEP), copy number variation (CNV), recombination signal sequence (RSS).

Grouping the adult ALL cases on the basis of molecular subtype reveals specific patterns in relation to the total number of deletions and amplifications (Supplementary Table 5). ALL cases have on average 8.15 deletions and 4.56 amplifications, which changes negligible when dichotomized on B-ALL, 8.06 deletions and 4.74 amplifications and T-ALL, 8.40 deletions and 4.10 amplifications.

Table 1. Clinical and molecular data of ALL and AML patients characterized on DNA mapping arrays.

Characteristics	All patients n=73	B-ALL patients n=53	T-ALL patients n=20	Characteristics	AML patients n=100
Age (years)				**Age (years)**	
median (range)	34 (17-69)	36 (17-69)	31 (18-65)	median (range)	43 (15-72)
Gender				**Gender**	
male	43(59%)	27	16	male	54 (54%)
female	30(41%)	26	4	female	46 (46%)
WBC count (x10^9/L)				**WBC count (x10^9/L)**	
median (range)	17 (1-375)	49 (1-375)	67(6-338)	median (range)	36 (1-234)
>30	15	15	0	**Blasts at diagnosis (before ficoll)**	
>100	4	0	4	median (range)	73 (2-96)
Blasts at diagnosis (before ficoll)				**FAB classification**	
Bone marrow (%)	93 (26-99)	83 (25-98)	78 (43-99)	Unknown	8 (8%)
Peripheral blood (%)	66 (10-94)	54 (1-97)	56 (8-93)	M1	5 (5%)
Immunophenotype				M2	30 (30%)
B-cell	49 (67%)	49 (92%)	0	M3	13 (13%)
T-cell	20 (27%)	0	20 (100%)	M4	23 (23%)
Biphenotypic	4 (6%)	4 (8%)	0	M5	17 (17%)
Cytogenetics				M6	1 (1%)
t(9;22) (*BCR-ABL*)	15 (21%)	15 (28%)	0	RAEB	1 (1%)
Hyperdiploid	9 (12%)	9 (17%)	0	RAEB-T	1 (1%)
t(1;19) (*E2A-PBX*)	1 (1%)	1 (2%)	0	**Cytogenetics**	
TEL-AML1	0	0	0	Favorable karyotype	14 (14%)
SIL-TAL1	4 (5%)	0	4 (20%)	11q23 abnormalities	2 (2%)
NUP214-ABL1	1 (1%)	0	1 (5%)	inv(3)/t(3;3)	6 (6%)
SET-NUP214	1 (1%)	0	1 (5%)	-7/7q	8 (8%)
11q23 abnormalities	5 (7%)	5 (9%)	0	Normal karyotype	55 (55%)
Normal karyotype	7 (10%)	5 (9%)	2 (10%)	Failure	8 (8%)
Failure	5 (7%)	3 (6%)	2 (10%)	**Molecular aberrations**	
				FLT3-ITD	31 (31%)
				NPM1c+	33 (33%)
				DNMT3A mutant	20 (20%)
				CEBPA double mutants	12 (12%)

Favorable karyotype comprises AML cases with the inv(16)(p13q22)/t(16;16)(p13;q22), t(8;21)(q22;q22) or t(15;17)(q22;q12) cytogenetic aberrations.

Table 2. Genetic aberrations detected in B-ALL, T-ALL and AML patients.

Gene	Aberration Type	B-cell ALL (n=53)	Focal B-Cell ALL (n=53)	T-cell ALL (n=20)	Focal T-cell ALL (n=20)	Total (n=73)	Focal total (n=73)	Aberration	Aberration type	AML (n=100)	Focal AML (n=100)
CDKN2A/2B	Deletion	30 (57%)	19 (36%)	18 (90%)	13 (65%)	48 (66%)	32 (44%)	**-7/-7q**	(Partial) loss of chromosome	12 (12%)	0 (0%)
IKZF1	Deletion	25 (47%)	16 (30%)	1 (5%)	1 (5%)	26 (36%)	17 (23%)	**+13**	Gain of chromosome	5 (5%)	0 (0%)
PAX5	Deletion	19 (36%)	8 (15%)	6 (30%)	0 (0%)	25 (34%)	8 (11%)	**+11**	Gain of chromosome	4 (4%)	0 (0%)
RB1	Deletion	8 (15%)	4 (8%)	3 (15%)	3 (15%)	11 (15%)	7 (10%)	**+8**	Gain of chromosome	4 (4%)	0 (0%)
BTG1	Deletion	10 (19%)	9 (17%)	1 (5%)	0 (0%)	11 (15%)	9 (12%)	**-9q**	(Partial) loss of chromosome	2 (2%)	0 (0%)
KRAS	Deletion	5 (9%)	4 (8%)	1 (5%)	0 (0%)	6 (8%)	4 (5%)	***NF1/SUZ12***	Deletion	5 (5%)	5 (5%)
BTLA	Deletion	7 (13%)	4 (8%)	0 (0%)	0 (0%)	7 (10%)	4 (5%)	***BCR***	Gain	3 (3%)	3 (3%)
MKKS	Deletion	7 (13%)	5 (9%)	0 (0%)	0 (0%)	7 (10%)	5 (7%)	***MYH11***	Deletion	2 (2%)	2 (2%)
CRLF2	Activation by deletion	12 (23%)	3 (6%)	1 (5%)	1 (5%)	13 (18%)	4 (5%)	***MLL***	Gain	1 (1%)	1 (1%)
NF1/SUZ12	Deletion	5 (9%)	1 (2%)	3 (15%)	3 (15%)	8 (11%)	4 (5%)	***BCL11B***	Gain	2 (2%)	2 (2%)
TAL1	Deletion	0 (0%)	0 (0%)	5 (25%)	4 (20%)	5 (7%)	4 (5%)				
WT1	Deletion	0 (0%)	0 (0%)	3 (15%)	2 (10%)	3 (4%)	2 (3%)				

Aberration type, e.g., deletion or amplification, is specified for each genetic lesion and enumerated for B-ALL, T-ALL and AML. Focal genetic lesions are observed if the aberration is small and specifically affects the gene listed.

There is a substantial lower number of deletions in ALL cases with a *MLL*-translocation, while hyperdiploid cases are characterized with an increased number of amplifications due to the additional chromosomes. Strikingly, we observe a substantial higher number of deletions (17.50) in patients with genetic alterations resulting in the overexpression of the receptor *CRLF2*[34], recently demonstrated to frequently coincide with *JAK2* abnormalities[35] in BCR-ABL1-like ALL cases.[36] Finally, the subtypes encompassing BCR-ABL1, SIL-TAL and hyperdiploid ALL have similar numbers of deletions and amplifications with respect to ALL on average.

CNV analysis of AML cases revealed only a few specific recurrent large genetic aberrations comprising the loss of chromosome 7 or 7q, gain of chromosome 13, gain of chromosome 11, and the gain of chromosome 8. This analysis also revealed the focal deletion of the locus encompassing the genes *NF1* and *SUZ12* in 5 AML cases (Table 2, Supplementary Table 4), as previously reported in paediatric AML.[37] Interestingly, this deletion was also observed in 3 T-ALL cases highlighting that *NF1* inactivation might play a role in both myeloid and lymphoblastic adult leukemia.[37]

The *NF1* microdeletion perturbs both *NF1* and *SUZ12*

The *NF1* microdeletion is the only common genetic aberration observed in ALL and AML, including AML cases with acquired *NPM1* mutations or the cytogenetic abnormality t(16;16)(p13;q22) (Supplementary Table 6), and involves the deletion of a small region of chromosome 17, i.e., del(17)(q11.2). We normalized the gene expression profiles to ascertain if this deletion confers a specific gene expression signature (Figure 2). Empirically, the commonly deleted region encompasses the genes encoding for NF1 and SUZ12 (Figure 3), both known to play a pivotal role in the malignant transformation of cancer.[37-41] Recent NGS efforts demonstrated that both genes are perturbed in a multitude of cancers, due to microdeletions or loss of function mutations[38,41], culminating into the hypothesis that PRC2 loss cooperates with the activation of the RAS pathway. Mutations in the remaining wild type allele of *NF1* or *SUZ12* could exacerbate the phenotype due to the complete loss of a functional protein. Previous studies demonstrated that most mutations in *SUZ12* are located in the VEFS-box domain[40,41], important for the interaction with the polycomb repressive complex 2 (PRC2) catalytic subunit EZH2. Genomic characterization of exons 12-16, encompassing the *SUZ12* VEFS-box domain, through dHPLC and targeted deep sequencing in 230 unselected ALL patients, revealed somatic mutations in 3 T-ALL cases lacking the *NF1* microdeletion (Figure 4). Subsequent targeted sequencing of the complete coding sequences of *NF1* in patients with a *NF1* microdeletion revealed 1 AML and 1 T-ALL case with a complete loss of functional NF1 due to premature stop codon introducing mutations in the remaining wild type allele (Supplementary Table 6). In total we have 5 AML and 3 T-ALL cases with *NF1* microdeletions perturbing both *NF1* and *SUZ12*, and we identified 3 additional T-ALL cases with acquired mutations in *SUZ12*, including one case with a focal *EZH2* deletion.

Differential gene expression analysis of the normalized GEPs comparing acute leukemia cases with the *NF1* microdeletion (n=8) to those without (n=844) demonstrated that *NF1* and *SUZ12* are

the most strongly down regulated genes (Figure 5A-B, Supplementary Table 7), suggesting and underpinning the hypothesis that the RAS pathway is activated in concert with the loss of PRC2.[38]

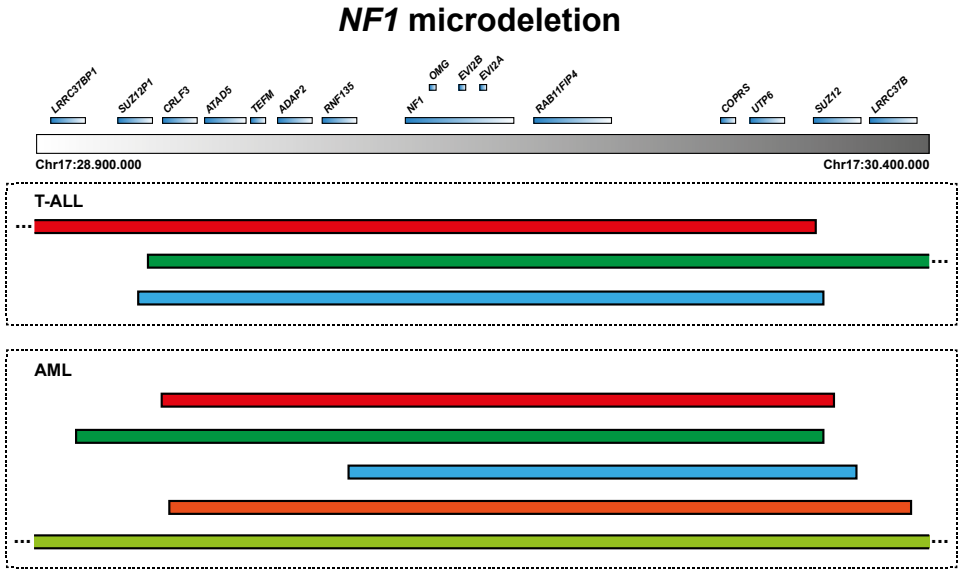


Figure 3. *NF1* microdeletion detected in T-ALL and AML. (top) The most commonly deleted genes due to the *NF1* microdeletion. (bottom) The colored bars represent the regions deleted for T-ALL and AML patients. The ends of the bar represent the breakpoint, while some patients have their breakpoint outside of the commonly deleted region. Note that all detected deletions affect the genes *NF1* and *SUZ12*.

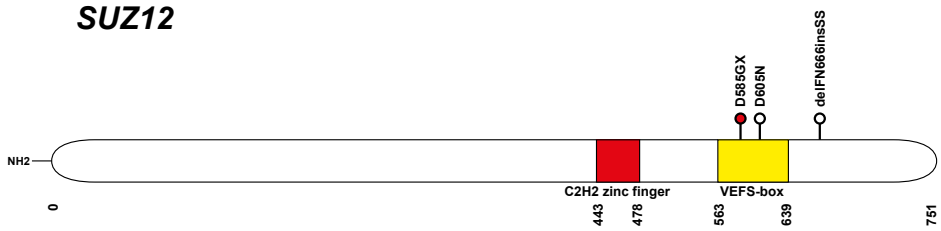


Figure 4. Somatic mutations detected in T-ALL. Screening a large cohort of T-ALL patients revealed somatic mutations, represented by the circles, in the gene encoding *SUZ12*. (red) premature stop codon, (open) substitution mutation.

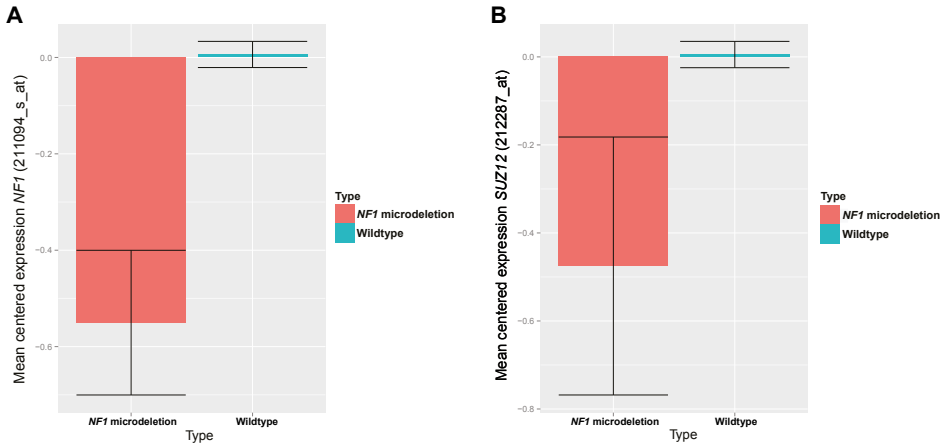


Figure 5. *NF1* microdeletion found in AML and T-ALL concurrently perturbs expression of *NF1* and *SUZ12*. (A) Comparison of *NF1* expression between acute leukemia patients with the deletion to all AML/ALL cases without reveals a substantial down regulation of *NF1*. (B) *SUZ12* expression levels are lower in *NF1* microdeletion acute leukemia cases as compared to all other AML/ALL cases.

Mutation and structural variation detection in 5 B-ALL cases

Copy number analysis revealed a substantial number of ALL cases with genetic lesions perturbing promoters or complete gene bodies of genes important in lymphoid development, postulated to be originating from illegitimate V(D)J recombinations.[8,9] To investigate whether this mechanism invokes the accumulation of genetic lesions in ALL we have performed WES and targeted resequencing of all breakpoint regions. We selected 5 B-ALL cases with substantial higher number of genetic lesions affecting promoters or genes involved in lymphoid development and determined all somatic mutations and structural variants. Of these 5 cases, 3 carry BCR-ABL1, 1 exhibits a BCR-ABL1-like GEP and 1 is a normal karyotype B-ALL case (Supplementary Table 8). In total 102 somatic mutations were detected, amounting to 20.4 somatic mutations per ALL case (Supplementary Table 9). No recurrent somatic mutation was detected precluding the identification of a commonly perturbed pathway on the basis of somatic mutation data. A complex *JAK2* mutation (I682>SP, Supplementary Figure 1) was detected in the BCR-ABL1-like patient with *CRLF2* overexpression, which is in line with previous studies.[36]

Structural variant analysis confirmed the copy number changes observed with DNA mapping arrays and the CNVsvd algorithm (Supplementary Table 10). In total 64 structural variants were detected, amounting to 12.8 structural variants per case on average. All cases had a complete or partial deletion of the *IKZF1* gene. We detected recurrent deletions of the promoters of *MKKS* (all cases), *BTLA* (3 cases), and *KRAS* (3 cases). Notably, other genes reported to be associated with ALL are recurrently deleted, i.e., *BTG1*, *PAX5*, *RB1*, and *CDKN2A/B*.

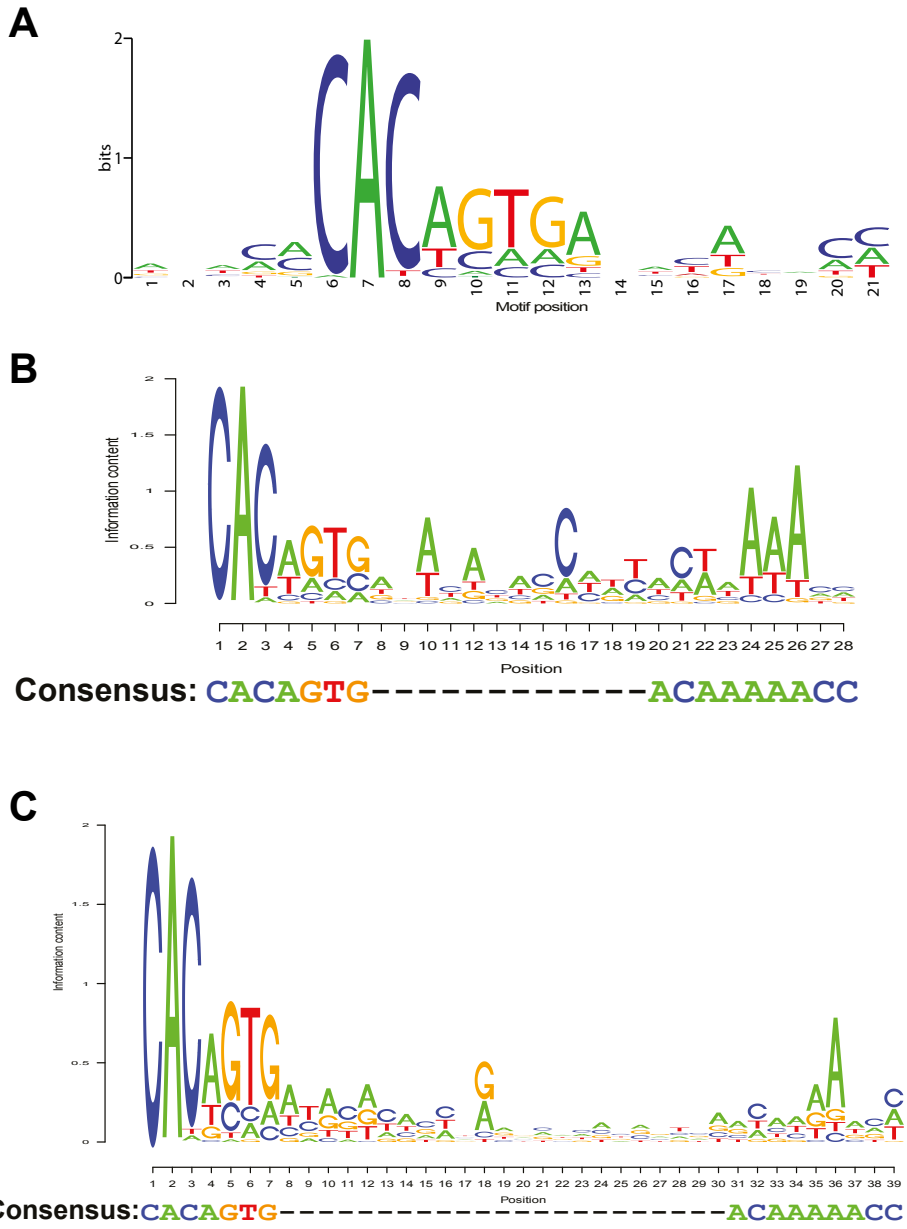


Figure 6. Consensus sequence of cRSS elements near B-ALL breakpoints. (A) *De novo* motif detected from the breakpoint sequences (B) Consensus sequence of cRSS elements with a 12 nucleotide spacer located near somatic deletion events. (C) Consensus sequence of cRSS with a 23 nucleotide spacer located near breakpoints.

Proximal cryptic RSS motifs imputes RAG activity

The 128 proximal breakpoint sequences of all the 64 structural variants were extracted and subsequently used for *de novo* motif detection with MEME.[29] This analysis discovered 1 significant motif characterized by the heptamer sequence CACAGTG (E-value $5.68X10^{-91}$, Figure 6A) present in 121 out of the 128 breakpoints. Strikingly, this motif equals the conserved heptamer sequence that is part of the RSS motif involved in V(D)J recombination. This process is mainly conferred by the RAG complex which targets V(D)J recombination sites by utilizing RSS motifs comprising of a highly conserved heptamer (CACAGTG), a 12-bp or 23-bp spacer without sequence specificity, and a less conserved nonamer (ACAAAAACC). This *de novo* analysis provides strong evidence that off-target RAG activity could mediate the deletions perturbing lymphoid specific genes.

To ascertain if the breakpoints boundaries are flanked by cryptic RSS motifs we uploaded the breakpoint sequences to RSSsite[28], which is designed to specifically detect cryptic RSS motifs, and demonstrated that most of the breakpoint boundaries are characterized by cryptic RSS motifs. Overall, in 58 out of the 64 rearrangements (90.6%), a confident RSS motif was found at one or both sides of the lesion (Supplementary Table 10). Strikingly, all *IKZF1* deletions in our 5 selected B-ALLs were characterized by a cryptic RSS motif proximal to the breakpoint on at least one side (Figure 7). All cryptic RSS motifs were extracted from RSSsite and used for the construction of a position weight matrix and sequence logo (Supplementary Table 11). As expected, we observed the 12-bp spacer RSS motif (Figure 6B) and the 23-bp spacer RSS motif (Figure 6C) with less conservation for the nonamer.

Further diversification of antigen loci is conferred by the amendment of palindromic sequences, through the opening of the recombination hairpins by the protein Artemis, or non-templated sequences randomly incorporated by terminal deoxynucleotidyl transferase (TdT).[42] We observed that 54 out of the 63 (85.7%) resolved rearrangements demonstrated the incorporation of non-templated sequences at the breakpoint (Supplementary Table 10), providing further evidence that RAG-mediated cleavage and further processing of the DNA ends play a pivotal role in the invocation of rearrangements in ALL.

Epigenetic state at RAG-induced structural variation boundaries

Most of the structural variants are characterized by a flanking cryptic RSS motif on one or both sides. We examined if the breakpoint junctions are enriched for particular epigenetic states or binding of known proteins. We procured the genome segmentation and Chip-Seq data of the B-lymphoblastic cell line GM12878 from ENCODE[31] due to its complete characterization. In total, we examined 125 breakpoint junctions (Supplementary Table 10), omitting 3 breakpoint junctions located in undetermined regions of the human genome sequence (hg19) and lacking ENCODE data for the updated human genome sequence (hg38). The genome segmentation divides the genome into 15 chromatin states based on different combinations of epigenetic markers. We observed a 39.9-fold enrichment of the breakpoints ($p < 2.2\ X10^{-16}$), within active promoter

regions (39 out of the 125 of the breakpoints) (Figure 8, Supplementary Table 12). Strikingly, the active promoter chromatin state is only assigned to 0.78% of the GM12878 genome. Additionally, we observed a strong enrichment for weak promoters (11.4-fold), poised promoters (19.3-fold), and strong enhancers (8.5-fold). Overall, this data implies that most breakpoints are not only characterized by cryptic RSS motifs, but also active epigenetic markers, e.g., H3K4me3.

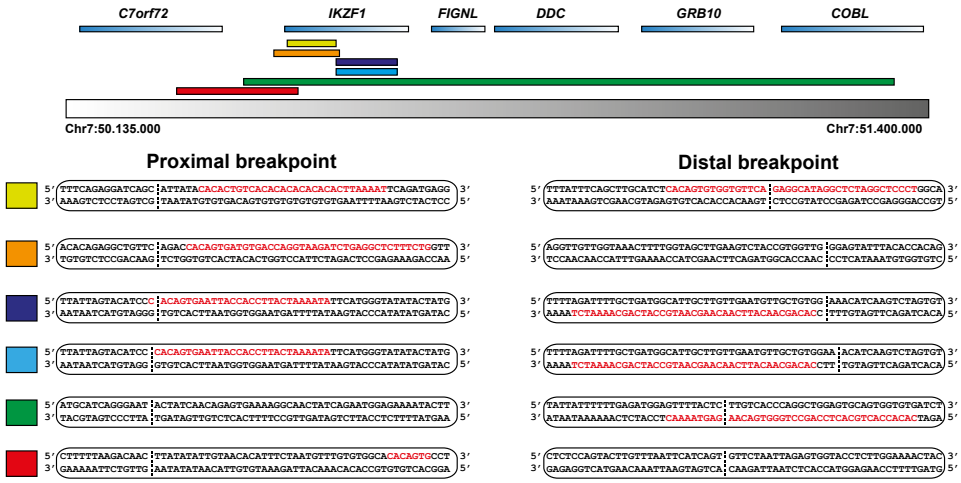


Figure 7. Detection of cryptic RSS sequences near the observed breakpoints affecting *IKZF1*. (top) Genes located in the region encompassing *IKZF1*. (middle) Colored bars illustrate the deletion events, perturbing *IKZF1*, detected in the 5 ALL patients. (bottom) Proximal and distal breakpoint sequences for every deletion event. The dashed line represents the breakpoint, while the red letters highlights the cryptic RSS detected by RSSsite.

Previous studies have shown that recombination foci in the V(D)J loci are characterized by H3K4me3, H3 acetylation, and RNA polymerase II binding.[10,43,44] We explored if particular histone markers and the binding of relevant proteins were enriched at the breakpoint loci using data generated from the cell line GM12878. The breakpoint loci are strongly enriched for H3K4me3 and H3K27ac (Figure 9A-B), while it is completely devoid for the repressive epigenetic marker H3K27me3 (Figure 9C). Reminiscent of the V(D)J rearrangement foci in the antigen receptors the breakpoints are likewise enriched for the binding of RNA polymerase II (Figure 9D). The RAG2 protein is able to bind H3K4me3 through its plant homology domain (PHD) and has been shown to bind to multiple regions outside of the antigen receptors in murine thymocytes.[10] We extracted Rag2 Chip-Seq data derived from murine thymocytes and translated the breakpoints positions detected in the B-ALL cases to homologous mouse genome positions. From the 125 breakpoint positions 65 could be translated to the murine genome. Strikingly, we observe a strong enrichment of Rag2 binding in murine thymocytes at the breakpoint loci (Figure 9E).

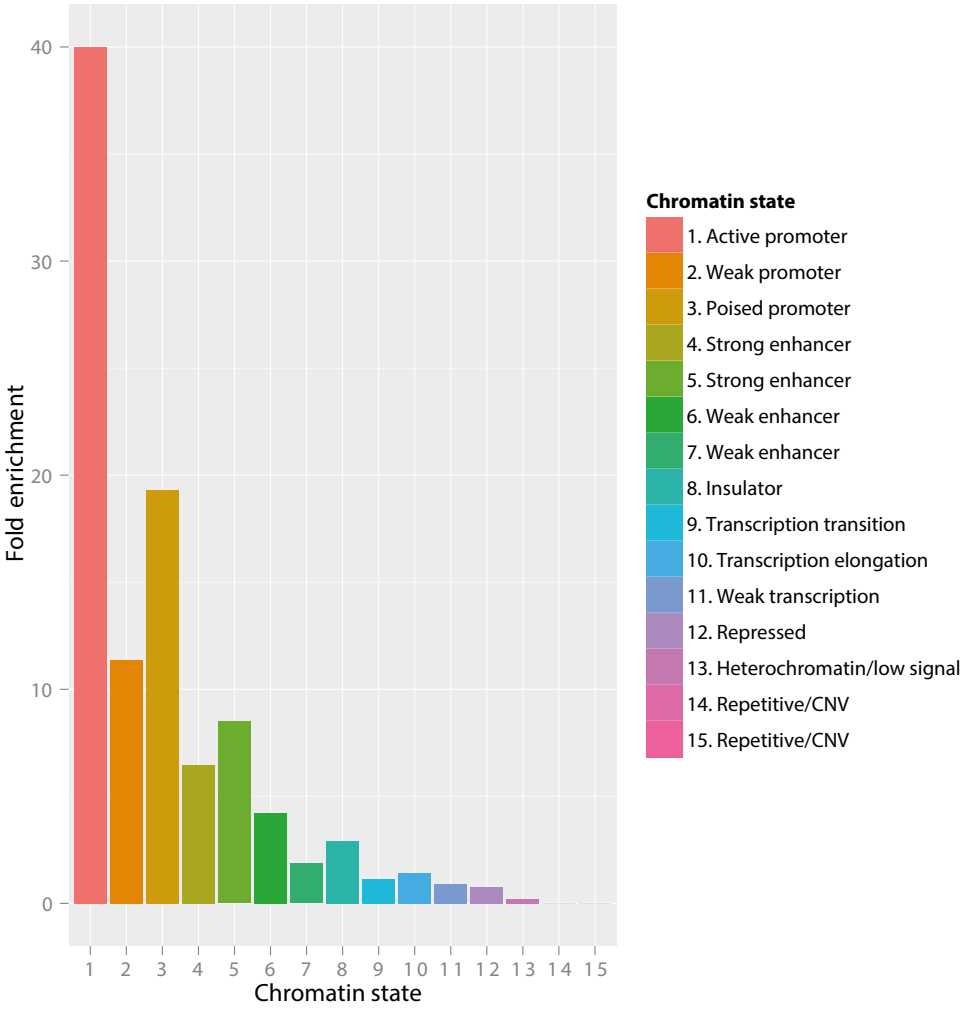


Figure 8. Chromatin state distribution of somatic structural variants. Somatic structural variations are segmented according to 1 of the 15 chromatin states extracted from ENCODE project data derived from the lymphoblastoid cell line GM12878. The structural variants are mainly located in active promoter regions which constitute only a small fraction, i.e., 0.78%, of the human genome.

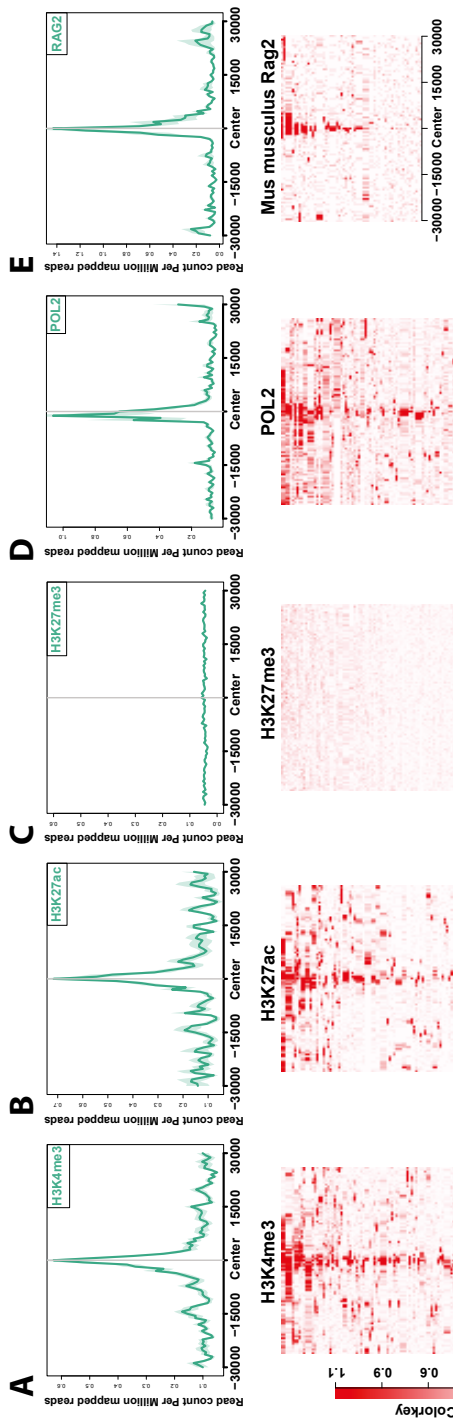


Figure 9. Epigenetic characterization of observed deletion breakpoint loci. Deletion breakpoint loci are enriched for active and open chromatin markers in the lymphoblastoid cell line GM12878 and Rag2 binding in homologous murine genomic locations (A) GM12878 H3K4me3 (B) GM12878 H3K27ac (C) GM12878 H3K27me3 (D) GM12878 RNA polymerase II (E) Rag2 binding in murine wild type thymocytes to loci homologous to the observed human B-ALL breakpoints.

Illegitimate RAG-mediated mutagenesis of RSS motifs in promoters and gene bodies

Exhaustive detection of complex insertion and deletion mutations revealed a substantial number of complex mutations situated within gene promoters and gene bodies. Careful inspection revealed that most of the complex insertions and deletions take place within cryptic RSS motifs, as determined by RSSsite, providing evidence of frequent open-and-shut joints outside antigen receptor loci invoked by the RAG complex.[45] Interestingly, we detected open-and-shut events in 3 out of the 5 B-ALL cases within a cryptic RSS motif located in the core promoter of *BTLA*. Strikingly, 2 out of the 3 B-ALL cases also lost the other allele due to a large genomic deletion (Figure 10). dHPLC analysis of the *BTLA* promoter revealed 8 additional B-ALL cases, predominantly (6 out of 8 cases) belonging to the BCR-ABL1 and BCR-ABL1-like subgroups, with open-and-shut joint events in the cryptic RSS motif (Supplementary Table 14). We also detected affected RSS motifs in the core promoter of *ADAR*, in conjunction with the loss of the other allele, and *LILRA2* (data not shown). Complex insertions and deletions in RSS motifs located within gene bodies have been observed in *TCF12*, *ARMC2, ZCCHC7, SCFD2,* and *PBX3* (Supplementary Figure 2). Finally, these complex mutations were also observed in RSS motifs located in classical V(D)J recombination foci within T-cell receptor and immunoglobulin genes.

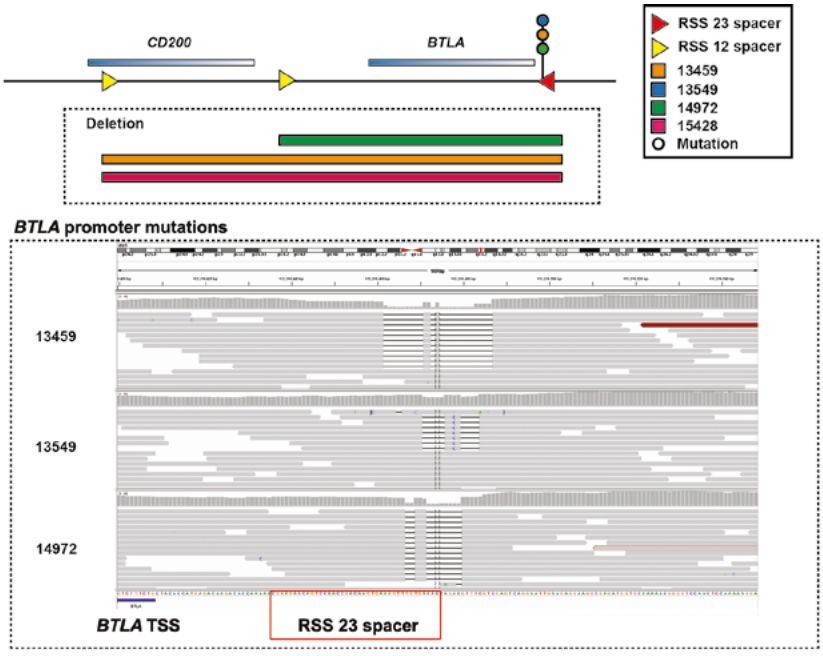


Figure 10. B-ALL somatic lesions in the promoter of *BTLA*. Somatic complex mutations have been detected in the promoters of developmental genes. For *BTLA* 3 B-ALL cases harbor complex mutations in a RSS motif within its promoter of which 2 B-ALL cases also have a deletion encompassing the complete gene.

DISCUSSION

This study provides the detailed genomic characterization of 50 B-ALL, 23 T-ALL, and 100 AML adult leukemia cases. We observed that many of the genetic lesions identified in paediatric ALL are likewise present in adult ALL. In comparison to array-based genome characterization studies performed on adult ALL[6] and paediatric ALL[5], we observed a substantial higher frequency of the most common loci perturbed by genetic lesions, e.g. *CDKN2A/B, IKZF1*, and *PAX5*. Strikingly, we demonstrated that all T-ALL cases acquired CNAs perturbing the CDKN2A/B pathway. An explanation of this divergence of genetic lesion frequency is the use of higher resolution DNA mapping arrays. In the case of the adult ALL cohort the Affymetrix 250K DNA mapping array was used[6], while for the paediatric cohort the Affymetrix 50K DNA mapping array.[5] We have utilized the Affymetrix 6.0 DNA mapping array which has approximately 8-fold and 20-fold more CNA measuring probe sets that the Affymetrix 250K and 50K DNA mapping array, respectively.

Interestingly, joint analysis of ALL and AML revealed a recurrent deletion affecting both the genes encoding for NF1 and SUZ12 in 3 T-ALL and 5 AML cases. The *NF1* gene encodes for neurofibromin 1 and is postulated to be a negative regulator of the RAS signal transduction pathway.[46] Loss of function mutations in the *NF1* locus results in hereditary neurofibromatosis type I[47], juvenile myelomonocytic leukemia[39], and AML.[46] The gene encoding SUZ12, a pivotal subunit of the polycomb repressive complex 2 (PRC2), which mediates the trimethylation of H3K27 (H3K27me3)[48] resulting in a repressive epigenetic mark. Recent sequencing efforts have demonstrated that the *SUZ12* gene is frequently perturbed by loss of function mutations in a multitude of cancers, including T-ALL[40] and malignant peripheral nerve sheet tumors (MPNSTs).[41] Subsequent GEP analyses revealed that both genes are substantially down regulated underpinning the recently proposed hypothesis that PRC2 complex loss cooperates with RAS pathway activation in a multitude of cancers. Finally, sequencing efforts revealed mutations within the VEFS-box domain of SUZ12, necessary for the interaction with EZH2, in additional T-ALL cases. In 2 out of 11 leukemia cases with perturbations in *SUZ12*, i.e., mutations or deletions, we observed premature stop codon introducing mutations in the remaining *NF1* wild type allele, demonstrating that aberrations affecting *NF1* and *SUZ12* could cooperate in the pathogenesis of adult leukemia.

A previous study has demonstrated that RAG-mediated recombination is a prominent driver of rearrangements in *ETV6-RUNX1* ALL[9], a rarely observed fusion protein in adult ALL. Here, we report on 5 BCR-ABL1/BCR-ABL1-like B-ALL cases for which the genomic rearrangements are predominantly driven by RAG recombination. Initial mutational analysis provided no recurrent somatic mutations, precluding the identification of a common pathway associated with leukemogenesis. However, *de novo* motif detection of breakpoint sequences hinted towards the pre-eminence of RAG-mediated rearrangements since 90.6% of the structural variants harbored a cryptic 12-bp or 23-bp spacer RSS motif flanking one or both of the breakpoint positions. We examined the enrichment of epigenetic markers near breakpoint loci and demonstrated, likewise

to classical V(D)J recombination[10], that these breakpoints are enriched for H3K4me3, H3 acetylation and RNA polymerase II binding. Chip-Seq data revealed Rag2 binding near the identified breakpoints in murine thymocytes, suggesting that these regions could be predispositioned for RAG-mediated recombination.

Finally, we observed open-and-shut joining events in gene promoters and gene bodies involving RSS motifs which are likely introduced by illegitimate RAG activity. RAG-mediated deletion of one allele in conjunction with the open-and-shut mutagenic event on the remaining allele could result in the complete knockout of a gene. Recurrence of these open-and-shut events, e.g. the *BTLA* core promoter, suggests that these mutagenic events are associated with leukemogenesis. The complex mutational patterns hint towards the binding of the RAG complex, subsequent nicking, addition of non-templated nucleotides, but the failure to form the synaptic complex leading to error-prone non-homologous end joining (NHEJ). This mutational process requires additional studies to elucidate the exact mechanism and it would therefore be interesting to replicate these results in additional ALL cases or subtypes.

The prospect of prominent RAG-mediated oncogenic rearrangements in adult leukemia provides an interesting topic for further research. Although, we have shown that it is a prominent mutational process, especially in regions characterized by RSS motifs and epigenetic markers native to V(D)J regions within antigen receptors, it is still unknown why the RAG recombinase invokes these lesions outside the antigen loci. Strikingly, most cases with a *BTLA* promoter mutation belong to the BCR-ABL1 or BCR-ABL1-like group. Previous studies have demonstrated that c-ABL1 and BCR-ABL1 modulate the activity and protein expression level of the NHEJ component DNA-dependent protein kinase catalytic subunit (DNA-PKcs).[49,50] Interestingly, a recent study has demonstrated that BCR-ABL1-like cases frequently acquire activating lesions in kinases[36] likewise to the BCR-ABL1 fusion product. However, if kinase-activating lesions modulate the NHEJ DNA repair pathway activity or behaviour remains a topic for further investigation.

Acknowledgments

We thank I.P. Touw, T. Cupedo and other members of the Department of Hematology at the Erasmus MC for their support. This work was financially supported by a grant from the Center for Translational Molecular Medicine (CTMM; GR03O-102, M. Sanders).

REFERENCES

1. Jabbour EJ, Faderl S, Kantarjian HM. Adult acute lymphoblastic leukemia. *Mayo Clin Proc*. 2005;80(11):1517-1527.
2. Faderl S, O'Brien S, Pui CH, et al. Adult acute lymphoblastic leukemia: concepts and strategies. *Cancer*. 2010;116(5):1165-1176.
3. Cancer Facts & Figures 2014. American Cancer Society. Available at http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf. Accessed September 15, 2014.
4. Mrozek K, Harper DP, Aplan PD. Cytogenetics and molecular genetics of acute lymphoblastic leukemia. *Hematol Oncol Clin North Am*. 2009;23(5):991-1010, v.
5. Mullighan CG, Goorha S, Radtke I, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007;446(7137):758-764.
6. Okamoto R, Ogawa S, Nowak D, et al. Genomic profiling of adult acute lymphoblastic leukemia by single nucleotide polymorphism oligonucleotide microarray and comparison to pediatric acute lymphoblastic leukemia. *Haematologica*. 2010;95(9):1481-1488.
7. Mullighan CG, Phillips LA, Su X, et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*. 2008;322(5906):1377-1380.
8. Waanders E, Scheijen B, van der Meer LT, et al. The origin and nature of tightly clustered BTG1 deletions in precursor B-cell acute lymphoblastic leukemia support a model of multiclonal evolution. *PLoS Genet*. 2012;8(2):e1002533.
9. Papaemmanuil E, Rapado I, Li Y, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet*. 2014;46(2):116-125.
10. Ji Y, Resch W, Corbett E, Yamane A, Casellas R, Schatz DG. The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. *Cell*. 2010;141(3):419-431.
11. Sanders MA, Valk PJ. The evolving molecular genetic landscape in acute myeloid leukaemia. *Curr Opin Hematol*. 2013;20(2):79-85.
12. Hahn CN, Chong CE, Carmichael CL, et al. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet*. 2011;43(10):1012-1017.
13. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med*. 2010;363(25):2424-2433.
14. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med*. 2009;361(11):1058-1066.
15. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365(15):1384-1395.
16. Kon A, Shih LY, Minamino M, et al. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat Genet*. 2013;45(10):1232-1237.
17. Radtke I, Mullighan CG, Ishii M, et al. Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proc Natl Acad Sci U S A*. 2009;106(31):12944-12949.
18. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*. 2009;94(1):131-134.
19. Wickham H. ggplot2: elegant graphics for data analysis: Springer; 2009.
20. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*. 2004;20(8):1233-1240.
21. Sanders MA, Verhaak RG, Geertsma-Kleinekoort WM, et al. SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels. *BMC Genomics*. 2008;9:41.
22. Klijn C, Holstege H, de Ridder J, et al. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res*. 2008;36(2):e13.
23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
24. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
25. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.

26. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865-2871.
27. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6(9):677-681.
28. Merelli I, Guffanti A, Fabbri M, et al. RSSsite: a reference database and prediction tool for the identification of cryptic Recombination Signal Sequences in human and murine genomes. *Nucleic Acids Res*. 2010;38(Web Server issue):W262-267.
29. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 1994;2:28-36.
30. Bembom O. seqLogo: Sequence logos for DNA sequence alignments. R package version 1.32.1. 2014.
31. ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
32. Shen L, Shao N, Liu X, Nestler E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*. 2014;15:284.
33. Usvasalo A, Savola S, Raty R, et al. CDKN2A deletions in acute lymphoblastic leukemia of adolescents and young adults: an array CGH study. *Leuk Res*. 2008;32(8):1228-1235.
34. Mullighan CG, Collins-Underwood JR, Phillips LA, et al. Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. *Nat Genet*. 2009;41(11):1243-1246.
35. Hertzberg L, Vendramini E, Ganmore I, et al. Down syndrome acute lymphoblastic leukemia, a highly heterogeneous disease in which aberrant expression of CRLF2 is associated with mutated JAK2: a report from the International BFM Study Group. *Blood*. 2010;115(5):1006-1017.
36. Roberts KG, Li Y, Payne-Turner D, et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N Engl J Med*. 2014;371(11):1005-1015.
37. Balgobind BV, Van Vlierberghe P, van den Ouweland AM, et al. Leukemia-associated NF1 inactivation in patients with pediatric T-ALL and AML lacking evidence for neurofibromatosis. *Blood*. 2008;111(8):4322-4328.
38. De Raedt T, Beert E, Pasmant E, et al. PRC2 loss amplifies Ras-driven transcription and confers sensitivity to BRD4-based therapies. *Nature*. 2014.
39. Side LE, Emanuel PD, Taylor B, et al. Mutations of the NF1 gene in children with juvenile myelomonocytic leukemia without clinical evidence of neurofibromatosis, type 1. *Blood*. 1998;92(1):267-272.
40. Ntziachristos P, Tsirigos A, Van Vlierberghe P, et al. Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat Med*. 2012;18(2):298-301.
41. Lee W, Teckie S, Wiesner T, et al. PRC2 is recurrently inactivated through EED or SUZ12 loss in malignant peripheral nerve sheath tumors. *Nat Genet*. 2014.
42. Gellert M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu Rev Biochem*. 2002;71:101-132.
43. Shimazaki N, Tsai AG, Lieber MR. H3K4me3 stimulates the V(D)J RAG complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations. *Mol Cell*. 2009;34(5):535-544.
44. Karo JM, Schatz DG, Sun JC. The RAG Recombinase Dictates Functional Heterogeneity and Cellular Fitness in Natural Killer Cells. *Cell*. 2014;159(1):94-107.
45. Lewis SM, Hesse JE. Cutting and closing without recombination in V(D)J joining. *The EMBO journal*. 1991;10(12):3631.
46. Bollag G, Clapp DW, Shih S, et al. Loss of NF1 results in activation of the Ras signaling pathway and leads to aberrant growth in haematopoietic cells. *Nat Genet*. 1996;12(2):144-148.
47. Wallace MR, Marchuk DA, Andersen LB, et al. Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. *Science*. 1990;249(4965):181-186.
48. Hansen KH, Bracken AP, Pasini D, et al. A model for transmission of the H3K27me3 epigenetic mark. *Nat Cell Biol*. 2008;10(11):1291-1300.
49. Kharbanda S, Pandey P, Jin S, et al. Functional interaction between DNA-PK and c-Abl in response to DNA damage. 1997.
50. Deutsch E, Dugray A, AbdulKarim B, et al. BCR-ABL down-regulates the DNA repair protein DNA-PKcs. *Blood*. 2001;97(7):2084-2090.

CHAPTER 6

Sequential gain of mutations in severe congenital neutropenia progressing to acute myeloid leukemia

Renée Beekman[1], Marijke Valkhof[1], Mathijs A. Sanders[1], Paulette van Strien[1], Jurgen R. Haanstra[1], Lianne Broeders[1], Wendy M. Geertsma-Kleinekoort[1], Anjo J.P. Veerman[2], Peter J.M. Valk[1], Roel G. Verhaak[3], Bob Löwenberg[1], Ivo P. Touw[1]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands

[2] Vrije Universiteit Medical Center, Department of Pediatric Hematology-Oncology, Amsterdam, The Netherlands

[3] MD Anderson Cancer Center, Department of Bioinformatics and Computational Biology, Houston TX, USA

Blood 2012 May 31;119(22):5071-7

ABSTRACT

Severe congenital neutropenia (SCN) is a bone marrow failure syndrome with a high risk to progress towards acute myeloid leukemia (AML). The underlying genetic changes involved in SCN evolution to AML are largely unknown. We obtained serial hematopoietic samples of an SCN patient who developed AML 17 years after initiation of granulocyte-colony stimulating factor (G-CSF) treatment. Next-generation sequencing was done to identify mutations during disease progression. In the AML phase, we found 12 acquired non-synonymous mutations. Three of these, in *CSF3R*, *LLGL2* and *ZC3H18*, co-occurred in a subpopulation of progenitor cells already in the early SCN phase. This population expanded in time, whereas clones solely harboring *CSF3R* mutations disappeared from the bone marrow. The other 9 mutations were only apparent in the AML phase and affected known AML-associated genes *(RUNX1* and *ASXL1*) and chromatin remodelers (*SUZ12* and *EP300)*. In addition, a novel *CSF3R* mutation was found conferring autonomous proliferation to myeloid progenitors. We conclude that progression from SCN towards AML is a multistep process with distinct mutations arising early during the SCN phase and others later in AML development. Sequential gain of two *CSF3R* mutations implicates abnormal G-CSF signaling as a driver of leukemic transformation in this case of SCN.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter6/

INTRODUCTION

Severe congenital neutropenia (SCN) is a bone marrow failure syndrome characterized by strongly reduced neutrophil counts and recurrent, potentially life threatening, opportunistic bacterial infections. Treatment with granulocyte-colony stimulating factor (G-CSF) elevates peripheral neutrophil counts and reduces the risk of infections.[1] Leukemic progression of SCN is a major concern with an estimated overall cumulative incidence of approximately 20% after 15 years of G-CSF treatment.[2]

Constitutional mutations in the gene encoding neutrophil elastase (*ELANE*) are common defects in SCN.[3] In addition, the acquisition of nonsense mutations in the gene encoding the granulocyte-colony stimulating factor receptor (*CSF3R*) is a unique feature in SCN patients.[4-7] These mutations lead to expression of truncated CSF3R proteins, also known as delta forms. In cell line models, truncated CSF3R proteins are hampered in transducing signals required for proper neutrophil differentiation. Additionally, they confer increased proliferative responses to G-CSF treatment but do not cause leukemia in mice.[4-6,8-11] *CSF3R* delta mutations can be detected in approximately 30% of SCN patients. In some cases, distinct clones with different *CSF3R* delta mutations are present for many years.[7,12] After evolution of SCN towards AML, *CSF3R* delta mutations are found in approximately 80% of the cases.[12] Until now, all reported SCN/AML cases harboring a *CSF3R* delta mutation in the SCN phase also carry this mutation in the leukemic phase. These observations suggest that leukemic progression in SCN follows a unique pattern, with *CSF3R* delta mutations as an early event, followed by additional genetic and epigenetic events that are essential for full leukemic transformation. Chromosomal aberrations, such as loss of chromosome 7 and gain of chromosome 21, are apparent in AML arising from SCN and other bone marrow failure syndromes like Fanconi anemia and Shwachman-Diamond syndrome.[13] However, mutations that are quite commonly seen in *de novo* AML have not been reported in AML arising from SCN.[14] Thus, the additional molecular events involved in leukemic progression of SCN remain largely unknown.

To identify the sequential genetic events in leukemic progression of SCN towards AML, we collected serial hematopoietic samples of an SCN patient who developed AML after 17 years of G-CSF therapy. Using whole exome sequencing, we found 12 somatic non-synonymous mutations in the leukemic blasts of this patient. Three of these mutations, the known *CSF3R* mutation and mutations in *LLGL2* and *ZC3H18*, were already present at low frequencies in the early SCN phase, 15 years before AML was diagnosed. Myeloid colony analysis showed that these 3 "early" mutations co-existed in the same hematopoietic progenitors in a small subpopulation of bone marrow cells. Six years later, in the "intermediate" SCN phase, still 9 years before the AML became overt, we observed an expansion of the clone harboring all 3 mutations. The other 9 mutations were only apparent in the AML. The latter "late" appearing mutations comprise a second, novel, *CSF3R* mutation in addition to a series of novel and known AML-associated mutations. The novel *CSF3R* mutation is located on the already mutated *CSF3R-d715* allele and causes growth factor independence of myeloid progenitors.

MATERIAL AND METHODS

Case report

A 27-year old male SCN patient was diagnosed with AML 17 years after the start of G-CSF treatment (10µg/kg/day), on which he reached normal neutrophil counts. The patient had a constitutional heterozygous *ELANE* mutation, G174R. At the age of 12, 2 years after G-CSF treatment was initiated, a *CSF3R* delta mutation (*CSF3R-d715*) was discovered in the bone marrow.[6] At the time of AML diagnosis, the peripheral blood contained 24% blasts and dysplasia was observed in the bone marrow. G-CSF treatment was stopped at this point. Six weeks later, a bone marrow analysis revealed 17% blasts. Immunophenotypically, these blasts were of myeloid origin, i.e., positive for CD34, CD117, CD13, CD133, CD33, MPO and CD90. Because no HLA-identical donor was available, the patient received a matched unrelated donor (MUD) allogeneic bone marrow transplantation. Induction therapy was given according to the induction therapy scheme HOVON42A of the Hemato-Oncology Foundation for Adults in the Netherlands.[15] At initiation of induction therapy, the bone marrow contained 15.7% blasts, with 10-50% dysplasia in all lineages. Routine cytogenetic and molecular diagnostics revealed a trisomy 21 (47, XY, +21 [14] /46, XY [4]), with no additional abnormalities (*AML-ETO, CBFB-MYH11, FLT3ITD, FLT3TKD*, mutations in *NPM1, NRAS, KRAS, c-KIT, JAK2* and *CEBPA*). After the second induction cycle trisomy 21, was undetectable in a marrow cytogenetic analysis. The MUD transplantation was administered after myeloablative conditioning with chemotherapy and total body irradiation. Two months after the transplantation 28% blast were detected in the bone marrow, indicating a recurrence of the AML and the patient died 3.5 months after the transplant. Figure 1 gives a schematic overview of the disease history.

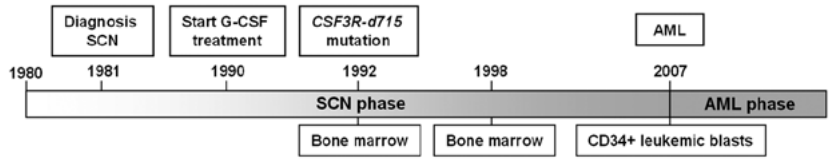


Figure 1. Chronological overview of the clinical course of the SCN/AML patient. Distinct events in the disease course are indicated above the timeline, i.e., the diagnosis of SCN, the initiation of G-CSF therapy, the discovery of the *CSF3R*-d715 mutation and the diagnosis of AML.

Patient cell samples

Ficoll-gradient separated bone marrow cells from the SCN phases and CD34+ leukemic blasts from the peripheral blood in the leukemic phase were used. Control DNA was isolated from bone marrow-derived fibroblasts. All cell samples were obtained and frozen according to established procedures for viable cell cryopreservation as previously described.[16] The study was performed under the permission of the Institutional Review Board of the Erasmus MC, registration number MEC-2008-387 for biobanking and MEC-2012-030 for the genetic analysis of leukemic progression in SCN patients.

Nucleotide sequencing

Whole exome sequencing (WES).

Sequencing libraries were prepared according to the SureSelect Target Enrichment system for Illumina, protocol version 2.2.1, Nov. 2010. In short, 3 µg genomic DNA was sheared to fragments of approximately 170 base pairs using the Covaris S-series Single Tube Sample Preparation System, Model S2 (Covaris, Woburn, MA, USA). Fragment sizes were checked on the Bioanalyzer (Agilent, Santa Clara, CA). Adapter ligated libraries were prepared according to the manufacturer's protocol using the Paired-End Genomic DNA Sample Prep Kit PE-102-1001 (Illumina, San Diego, CA); 5 cycles of amplification were used. Five hundred ng of prepped library was taken for hybridization with the SureSelect Human All Exon Kit (G3362A, Agilent). A sample concentration of 5.5 picomolar was loaded for sequencing on the Hiseq2000 (Illumina) using 101-bp paired-end reads.

Sequencing reads were processed with the Casava pipeline (version 1.7, Illumina). For alignment the hg18/NCBI36 assembly (March 2006) was used. Detection of single nucleotide variants, deletions and insertions was performed with otherwise default settings, while snpCovCutoff and indelsCovCutoff were switched off. Variations detected in the AML sample in 2 independent sequence runs were further analyzed after removal of germline variations (present in the fibroblasts) and single nucleotide polymorphisms (SNPs, dbSNP).[17] Next, non-synonymous variants were determined. Integrative Genome Browser was used for sequence read visualization.[18]

Sanger sequencing.

WES results were validated by Sanger sequencing, performed according to the manufacturer's protocol (Applied Biosystems, Foster City, CA, USA) using primers indicated in Table S1. Before amplicon generation, genomic DNA or cDNA was first amplified using a Whole Genome Amplification kit (WGA2, Sigma-Aldrich, Zwijndrecht, The Netherlands). DNA was purified with a PCR purification kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol and diluted to 50 ng/µl. Hundred nanograms of amplified DNA was used for amplicon generation; cycling conditions were 30" at 95°C, 30" at the indicated annealing temperature (Table S1) and 45" at 72°C for 35 cycles. In some instances, the unamplified material was used directly for Sanger sequencing (Table S1).

Amplicon-based deep sequencing.

Amplicons were generated and purified according to the Amplicon Library Preparation Method Manual (version May 2010, Roche, Basel, Switzerland). Primers and annealing temperatures are indicated in Table S2; 35 cycles were used for amplification. DNA enriched beads, carrying the amplification products, were generated according to the emPCR Amplification Method Manual – Lib-A (version May 2010, Roche); a beads to amplicon ratio of 1:2 was used. Amplicons were

analyzed with the GS junior (Roche). Sequence reads were analyzed using the GS Amplicon Variant Analyzer (Roche). For the SCN samples, coverage of at least 1600 was achieved to identify mutations present in minor clones within the bone marrow. For the AML sample coverage of 80 was considered sufficient to validate mutations.

Human myeloid colony assay

Bone marrow was thawed at 37°C, washed twice with IMDM (Gibco Invitrogen, San Diego, CA) with 10% FCS (PAA laboratories, Pasching, Austria). Per 4 ml of culture medium, 2.9 ml MethoCult (H4230, Stem Cell Technologies, Vancouver, Canada), 980 µl IMDM and human GM-CSF (Immunex, Seattle, WA), human G-CSF (Neupogen, Amgen, Thousand Oaks, CA) and human IL-3 (R&D Systems, Minneapolis, USA) in final concentrations of respectively 2 ng/ml, 200 ng/ml and 25 ng/ml were used. Cells were plated at a density of 0.8 x 10^5/ml. After 2 weeks genomic DNA of single colonies was isolated, followed by amplification using the Whole Genome Amplification kit and Sanger sequencing of *CSF3R-d715*, *LLGL2* and *ZC3H18,* as described above. Results were validated in an independent round of whole genome amplification for (1) colonies harboring a mutation, (2) colonies with unclear sequences and (3) a number of randomly chosen non-mutated colonies to rule out amplification artifacts. All colonies harboring mutations in *CSF3R*, *LLGL2* or *ZC3H18* were also analyzed for the presence of the remaining 9 mutations found in the AML sample.

Murine colony assays

Four different CSF3R expression constructs (WT, *d715*, *T595I*, *d715/T595I*) were generated and retrovirally transduced into bone marrow cells of *Csf3r*-deficient FVB/N mice.[19] Colony assays of these transduced progenitors were performed as previously described.[20] Further details of these procedures are given in the Supplementary Materials and Methods.

RESULTS

Whole exome sequencing reveals acquired mutations in SCN/AML

WES was done on genomic DNA from the CD34+ leukemic blast fraction and the fibroblast control sample. Acquired non-synonymous mutations were detected by identification of single nucleotide variants and small insertions and deletions, followed by subtraction of variants present in the control fibroblasts and known single nucleotide polymorphisms.[17] Twelve non-synonymous acquired mutations were identified and validated by Sanger sequencing (Table 1, Figure S1). Except for the mutation in *FBXO18*, all mutations occurred in evolutionary conserved amino acids (Figure S2). With the exception of *LAMB1*, all mutant transcripts were detectably expressed in the leukemic blasts (Figure S3). Mutations in *ASXL1* and *RUNX1* are known in myeloid malignancies.[21,22]

Table 1. Somatic non-synonymous mutations in SCN/AML.

Gene Symbol	RefSeq Reference Transcript	Genomic DNA Change (NCBI36/hg18)	cDNA Change	Mutation Type	Amino Acid Change	Protein Change
ASXL1	NM_015338.5	g.chr20:30485948dupA	c.1772dupA	Indel frameshift	Frameshift and premature stop	p.Y591*
CCDC155	NM_144688.4	g.chr19:54601976C>T	c.820C>T	Missense	Arg>Trp	p.R274W
CSF3R-T595I	NM_000760.3	g.chr1:36706021G>A	c.1853C>T	Missense	Thr>Ile	[$]p.T595I
CSF3R-d715	NM_000760.3	g.chr1:36704841G>A	c.2215C>T	Nonsense	Gln>*	[$]p.Q716*
EP300	NM_001429.3	g.chr22:39902447_ 39902453delTGGAGAC	c.5030_ 5036delTGGAGAC	Indel frameshift	Frameshift and premature stop	p.V1677Dfs*30
FBXO18	NM_032807.3	g.chr10:6003435C>G	c.2372C>G	Missense	Ala>Gly	p.A791G
LAMB1	NM_002291.2	g.chr7:107387385delG	c.2445delC	Indel frameshift	Frameshift and premature stop	p.P815Pfs*65
LLGL2	NM_004524.2	g.chr17:71070826G>C	c.665G>C	Missense	Arg>Pro	p.R222P
MGA	NM_001164273.1	g.chr15:39787311C>T	c.2282C>T	Missense	Pro>Leu	p.P761L
RUNX1	NM_001754.4	g.chr21:35153662C>T	c.592G>A	Missense	Asp>Asn	p.D198N
SUZ12	NM_015355.2	g.chr17:27346889_ 27346891dupATT	c.1789_ 1791dupATT	Indel	Insertion Ile	p.597dupI
ZC3H18	NM_144604.3	g.chr16:87192175delC	c.777delC	Indel frameshift	Frameshift and premature stop	p.P259Pfs*15

All 12 somatic non-synonymous mutations identified in the AML phase are listed. For each mutation, Refseq reference transcripts, the position of the mutation on genomic DNA, cDNA and protein level, the mutation type and the effect on the protein are indicated. See also Figure S1-S3. [$]Amino acid numbers based on earlier publications.[4,6]

Deletions in *EP300*, distinct from the 7-bp deletion found in this patient, have been reported in lymphomas.[23,24] The ATT insertion in *SUZ12* duplicates an isoleucine at amino acid position 597, located in the conserved VEFS-box. Mutations in this region, which is involved in the interaction between SUZ12 and the histone methyltransferase EZH2 in the polycomb repressor complex 2 (PRC2), have recently also been identified in myelodysplastic/myeloproliferative neoplasms (MDS/MPN) with 17q abnormalities.[25] As expected, the previously identified *CSF3R* delta mutation (*CSF3R-d715)* was present in the leukemic blasts, but remarkably a new *CSF3R* mutation, *T595I*, was now also present. Furthermore, the *CSF3R-T595I* mutation was located on the same allele as the delta mutation, as determined by Sanger sequencing of single amplicons generated from cDNA. Using exome sequencing data from 199 AML cases reported by The Cancer Genome Atlas (TCGA), a similar single *CSF3R-T595I* mutation was detected. Additionally, mutations in *ASXL1* (n=5), *CCDC155* (n=1), *LLGL2* (n=1), *MGA* (n=1), *RUNX1* (n=17), *SUZ12* (n=2) and *ZC3H18* (n=2) were found in the TCGA data set (Table S3).[26]

Amplicon-based sequencing reveals an early pre-leukemic clone that expands over time

Amplicon-based deep sequencing was applied to analyze the presence of all 12 somatic mutations in the bone marrow samples obtained at 15 and 9 years before AML was diagnosed (Figure 1). Not only the known *CSF3R-d715* mutation, but also mutations in *LLGL2* and *ZC3H18* were already present in these earlier disease phases (Figure 2A, Table S4). We investigated the clonal hierarchy of these mutations in single myeloid colonies cultured from the earliest bone marrow sample (15 years before AML diagnosis). In the individual colonies (n=88), the mutation status of *CSF3R-d715*, *LLGL2* and *ZC3H18* was determined. Fifteen colonies (17%) harbored both the *CSF3R-d715* and the *LLGL2* mutation, whereas none of the colonies exhibited exclusively either the *LLGL2* or the *CSF3R-d715* mutation (Figure 2B, Table S5). Two of the *CSF3R-d715* and *LLGL2* mutated colonies also carried the *ZC3H18* mutation (Figure 2B, Table S5), indicating that this mutation had emerged later in time. None of the other 9 mutations found in the AML cells was apparent in these colonies (Table S5).

A previous report has shown that multiple *CSF3R* delta mutations can be present in distinct progenitors in the bone marrow of an individual SCN patient.[7] In line with this previous report, we found myeloid colonies with *CSF3R-d717* (n=2) and *CSF3R-d725* (n=1) (Figure 2B, Table S5). Each of these mutations and yet an additional delta mutation (*CSF3R-d730*) were detected in the SCN phase at low frequencies by amplicon-based deep sequencing (Figure 2C, Table S6). None of these variant *CSF3R* mutant clones harbored *LLGL2* or *ZC3H18* mutations, nor were they seen as dominant clones in the AML (Figure 2, Table S5-S6). No viably frozen cells were available from the bone marrow sample obtained 9 years before AML development and colony analysis could not be performed at this stage. However, by amplicon-based deep sequencing we observed a parallel increase of the *CSF3R-d715*, *LLGL2* and *ZC3H18* mutations from 15 to 9 years before AML

development (Figure 2A). Together with the finding that these mutations are present in the same myeloid progenitor cells (Figure 2B), this observation is consistent with a selective outgrowth of clones carrying these 3 mutations.

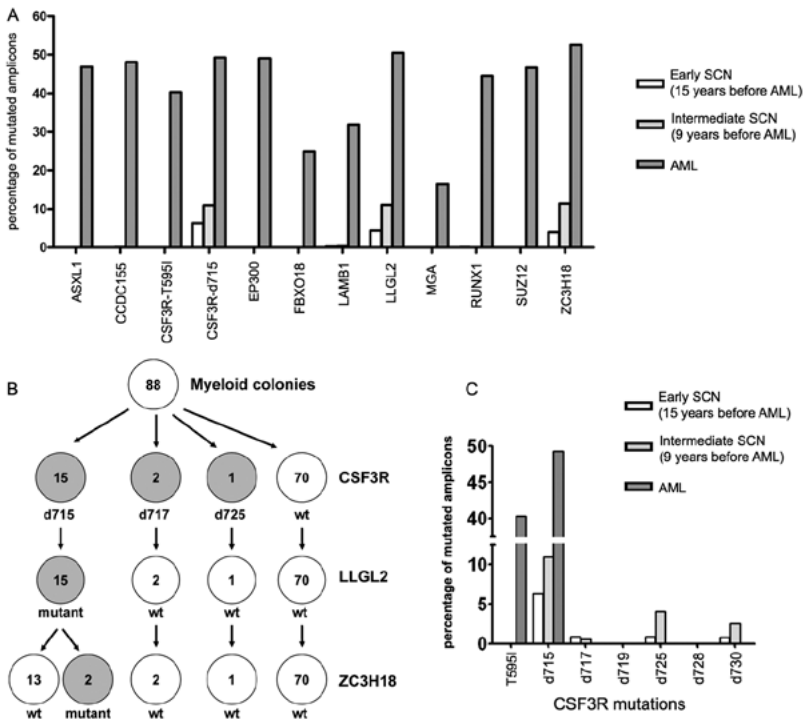


Figure 2. Acquisition of mutations in the evolution of SCN towards AML. (A) The 12 somatic non-synonymous mutations identified in the leukemic blasts were analysed in the SCN phase using amplicon-based deep sequencing. Per mutation, the percentage of mutated amplicons is shown. (B) Single myeloid colonies grown from the bone marrow sample obtained 15 years before leukemia development were analysed for the presence of mutations in *CSF3R*, *LLGL2* and *ZC3H18*. See also Table S5. (C) The presence of different *CSF3R* mutations in the bone marrow obtained 15 and 9 years before leukemia development and in the leukemic phase was investigated by amplicon-based deep sequencing. Per mutation, the percentage of mutated amplicons is shown. *T595I*: *CSF3R* mutation T595I, *d715-d730*: *CSF3R* delta mutations at amino acid position 715 to 730.

Sequential gain of a second *CSF3R* mutation results in G-CSF independence

A new *CSF3R* mutation, acquired at the *CSF3R-d715* mutant allele, was found exclusively in the AML blasts and changed a polar threonine residue at amino acid position 595 into a nonpolar isoleucine. This residue is located in a highly conserved threonine-rich region in the extracellular domain of the G-CSF receptor (Figure S2). Introduction of human *CSF3R* mutant receptors, carrying this new *T595I* mutation (Figure 3A), into Csf3r-deficient primary mouse bone marrow

progenitors resulted in the autonomous outgrowth of myeloid colony-forming cells (Figure 3, Table S7). Thus, in the AML phase of disease evolution two different co-existing mutations, i.e., the *T595I* single amino acid substitution and the *CSF3R-d715* mutation had accumulated in the gene encoding the G-CSF receptor. Because expression of the new *CSF3R* mutant without the delta mutation conferred G-CSF independence as did the mutant receptor carrying both the delta and the extracellular mutation, this gain of function can entirely be attributed to the *T595I* mutation. However, the *T595I/d715* colonies were bigger than the *T595I* colonies (Figure S4), which is suggestive of a higher proliferation capacity by the addition of the *CSF3R-d715* mutant.

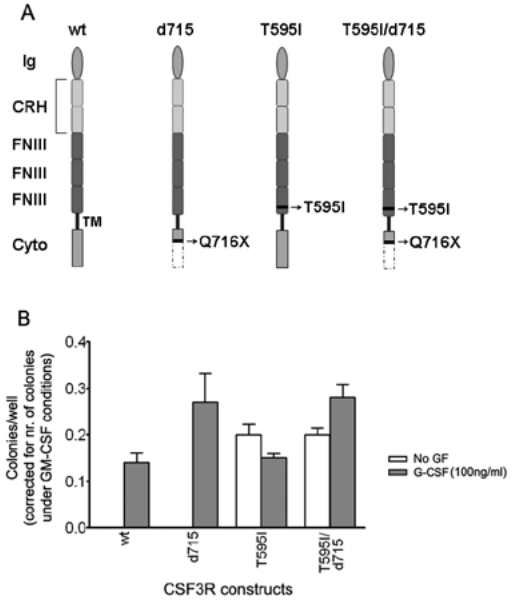


Figure 3. Functional analysis of *CSF3R* mutants in myeloid progenitor cell assays. In vitro colony growth of Csf3r-deficient murine hematopoietic progenitor cells expressing different *CSF3R* mutants. (A) Graphical representation of the different *CSF3R* constructs. Wild type (wt), T595I (containing the extracellular mutation at amino acid position 595), d715 (containing the intracellular mutation, Q716X, causing the introduction of a stop codon at amino acid position 716) and T595I/d715, containing both mutations as found in the SCN/AML patient. Ig: Immunoglobulin like domain; CRH: cytokine receptor homology domain; FNIII: fibronectin type III repeats; TM: transmembrane domain; cyto: cytoplasmic domain. Nomenclature has been adopted from Layton et al.[27] (B) Colonies were grown in the presence of puromycin, either without growth factor (no GF) or with 100ng/ml human G-CSF. The induced colony growth is dependent on the transduction efficiency and the type of *CSF3R* construct. The transduction efficiency can be deduced from the number of GM-CSF-induced colonies under puromycin selection as the *CSF3R* constructs confer puromycin resistance, but do not affect GM-CSF-induced colony growth. Hence, by dividing the number of colonies by the number of GM-CSF induced colonies the transduction efficiency was corrected for.

DISCUSSION

The results of the present study identified non-synonymous mutations acquired in an SCN patient who progressed to AML.The availability of sequential hematopoietic samples from the childhood SCN phase to overt AML, spanning a period of 17 years, provided the unique opportunity to identify the early and late genetic defects associated with leukemic progression (Figure 4). The *CSF3R-d715* mutation and a mutation in *LLGL2,* encoding the human homologue of the Drosophila lethal giant larvae (*Lgl*) gene, were the first 2 acquired mutations in the early SCN phase. Loss of *Lgl* in Drosophila leads to inadequate distribution of the cell polarity protein Numb, resulting in inappropriate cell fate determinations and tumor formation in epithelial tissues and the brain.[28-30] In man, the NUMB protein has been implicated in controlling the balance between symmetric versus asymmetric hematopoietic stem cell divisions. Interestingly, deregulation of NUMB expression has been associated with blast transformation of chronic myeloid leukemia.[31,32] How the *LLGL2* mutation found in this study affects hematopoietic stem cell divisions is still unknown; however, the fact that *CSF3R-d715* and *LLGL2* mutations were uniformly present in the same myeloid cells could suggest that they cooperate. Hierarchically, the next genetic abnormality occurring in the early SCN phase in the *CSF3R-d715*- and *LLGL2*-mutated clone was a mutation in *ZC3H18*. ZC3H18 is a putative mRNA binding protein with a still unknown function, but in trypanosomes it is shown to be essential for differentiation.[33]

Additionally, we found small subpopulations harboring distinct *CSF3R* delta mutations in the bone marrow at the early SCN stage. All these clones disappeared during the disease course, except the *CSF3R-d715* clone which evolved towards AML. The different *CSF3R* delta mutations cause expression of distinct truncated G-CSF receptors that all have similar consequences for signaling; they result in a sustained activation of signal transducer and activator of transcription 5 (STAT5).[8] STAT5 is a transcription factor, implicated in abnormal signaling responses of leukemic cells with mutated forms of the FLT3 receptor (FLT3-ITD) in AML and the BCR-ABL fusion protein in CML.[34,35] Furthermore, why one of these *CSF3R* delta mutant clones survived *in vivo* and progressed towards a fully transformed AML clone while the other *CSF3R* delta variants extinguished during disease development currently remains unexplained. However, it is conceivable that the additional mutations in *LLGL2* and *ZC3H18*, exclusively present in the *CSF3R-d715* clone, conferred a competitive growth advantage of this particular subclone representative of essential early steps in leukemic progression that cooperate with the aberrant signaling from the truncated G-CSF receptor.

Besides early genetic events, we found 9 mutations that occurred later in the process of leukemic transformation. Of particular interest is the novel *CSF3R* mutation (*T595I*), which appeared exclusively in the AML stage and imposed growth factor independence on an already functionally defective G-CSF receptor. A different mutation in the *CSF3R* transmembrane domain, *CSF3R-T617N,* with a similar downstream effect was previously found as a constitutive mutation in a family with hereditary chronic neutrophilia and as an acquired mutation in 2 AML patients. This mutation is suggested to cause ligand independent homodimerization and induces growth factor

independent proliferation and differentiation.[36,37] The major difference between the *T617N* and the *T595I* mutation in our patient is that the latter one is located on the already affected *CSF3R-d715* allele, which has been shown to cause increased proliferation and impaired differentiation in cell line and animal models[8,38,39] and which could explain the increase in colony size between the *T595I* mutant and the *T595I/d715* mutant. The acquisition of autonomous growth abilities by myeloid progenitor cells that already express a hyper-responsive G-CSF receptor mutant strongly suggests that perturbed G-CSF signaling was of vital importance for malignant transformation in this case of SCN. To our knowledge, this is the first example of a gain of 2 different mutations in the same receptor in the process of malignant transformation. An important, but still open question, is whether the administration of G-CSF to this patient had contributed to the acquisition of this additional mutation. Possibly, the continuous proliferative pressure imposed by G-CSF on clones carrying mutations in *CSF3R-d715* and *LLGL2* and later also in *ZC3H18* may have provided the context for the selection of a clone harboring this self-activating *CSF3R* mutation, pushing it to become an autonomously proliferating and dominant leukemic clone.

Abnormalities appearing in the AML phase included mutations in *ASXL1*, *SUZ12*, and *EP300*, genes encoding proteins involved in chromatin modification. Mutations in *ASXL1* have been reported previously in AML and are associated with an unfavorable prognosis.[40] SUZ12 is a member of the PRC2 complex that also contains EZH2, the histone methyl transferase responsible for the di- and tri-methylation of lysine 27 in the tail of histone 3 (H3K27), imposing a chromatin mark that represses gene expression. Mutations affecting *EZH2* and less frequently *SUZ12* have been detected in MDS/MPN patients.[25,41,42] In contrast, mutations in *EP300* and the highly related *CREBBP*, encoding histone acetyl transferases that act as transcriptional co-activators, have not yet been reported in myeloid malignancies but are the most frequent structural abnormalities in follicular lymphoma and diffuse large B cell lymphoma.[23,24] Mutations in *CCDC155*, encoding coiled-coil domain containing protein 155 with unknown function; *FBXO18*, encoding a DNA helicase involved in DNA repair and genomic integrity; *LAMB1,* encoding an extracellular matrix protein; and *MGA*, encoding a Max gene associated antagonist of Myc oncoproteins, all represent novel mutations with currently unknown functional significance.

Recurrence is an important criterion to discriminate driver from passenger mutations in the process of malignant transformation. Interestingly, mutations in *CCDC155, LLGL2, MGA* and *ZC3H18* were recently also reported by the TCGA consortium in a panel of AML patients (n=199), albeit at low frequencies.[26] Because frequencies of specific mutations have been shown to vary with the natural history of AML, e.g. *de novo* versus secondary to MDS/MPN or different bone marrow failure syndromes[14,43], it will be of interest to establish how often the newly identified genes are affected in distinct subtypes of secondary AML. Specifically, it will be important to determine whether *LLGL2*, *ZC3H18* or functionally related genes are more generally affected in bone marrow failure syndromes prone to progress to AML and to establish how these mutations contribute to malignant transformation in conjunction with cooperative gene defects.

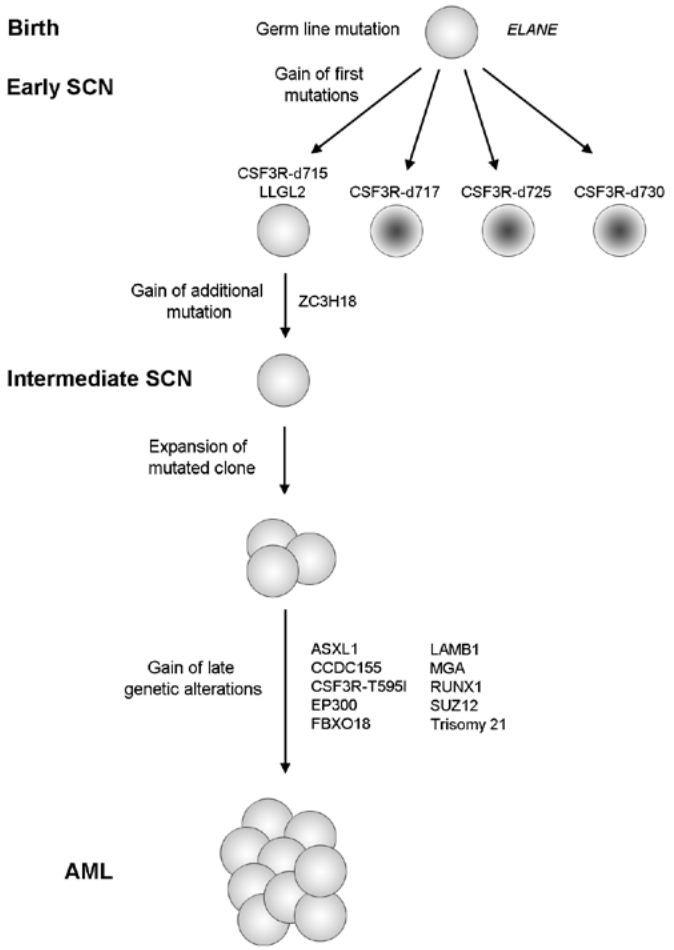


Figure 4. Schematic representation of the clonal evolution of SCN towards AML. The sequential genetic events, starting with the presence of a germ line mutation in the gene encoding neutrophil elastase (*ELANE)* are indicated. A sequential gain of *CSF3R* delta mutations and an *LLGL2* mutation is observed in the early SCN phase. Only the clone harboring the *CSF3R-d715* and the *LLGL2* mutation gained an additional mutation in *ZC3H18*, followed by its expansion in the intermediate SCN phase. Gain of 9 additional mutations and trisomy 21 in the mutated population preceded complete transformation towards AML. *CSF3R*-d715-d730: *CSF3R* delta mutations at amino acid position 715 to 730.

Acknowledgments

This research was supported by the Center of Translational Molecular Medicine (CTMM), The Dutch Cancer Society "KWF kankerbestrijding" and the E-RARE project ELA2-CN.

REFERENCES

1. Dale DC, Bonilla MA, Davis MW, et al. A randomized controlled phase III trial of recombinant human granulocyte colony-stimulating factor (filgrastim) for treatment of severe chronic neutropenia. *Blood*. 1993;81(10):2496-2502.
2. Rosenberg PS, Zeidler C, Bolyard AA, et al. Stable long-term risk of leukaemia in patients with severe congenital neutropenia maintained on G-CSF therapy. *Br J Haematol*. 2010;150(2):196-199.
3. Dale DC, Link DC. The many causes of severe congenital neutropenia. *N Engl J Med*. 2009;360(1):3-5.
4. Dong F, Brynes RK, Tidow N, Welte K, Lowenberg B, Touw IP. Mutations in the gene for the granulocyte colony-stimulating-factor receptor in patients with acute myeloid leukemia preceded by severe congenital neutropenia. *N Engl J Med*. 1995;333(8):487-493.
5. Dong F, Dale DC, Bonilla MA, et al. Mutations in the granulocyte colony-stimulating factor receptor gene in patients with severe congenital neutropenia. *Leukemia*. 1997;11(1):120-125.
6. Dong F, Hoefsloot LH, Schelen AM, et al. Identification of a nonsense mutation in the granulocyte-colony-stimulating factor receptor in severe congenital neutropenia. *Proc Natl Acad Sci U S A*. 1994;91(10):4480-4484.
7. Germeshausen M, Ballmaier M, Welte K. Incidence of CSF3R mutations in severe congenital neutropenia and relevance for leukemogenesis: Results of a long-term survey. *Blood*. 2007;109(1):93-99.
8. Hermans MH, Antonissen C, Ward AC, Mayen AE, Ploemacher RE, Touw IP. Sustained receptor activation and hyperproliferation in response to granulocyte colony-stimulating factor (G-CSF) in mice with a severe congenital neutropenia/acute myeloid leukemia-derived mutation in the G-CSF receptor gene. *J Exp Med*. 1999;189(4):683-692.
9. Touw IP, van de Geijn GJ. Granulocyte colony-stimulating factor and its receptor in normal myeloid cell development, leukemia and related blood cell disorders. *Front Biosci*. 2007;12:800-815.
10. Dong F, van Buitenen C, Pouwels K, Hoefsloot LH, Lowenberg B, Touw IP. Distinct cytoplasmic regions of the human granulocyte colony-stimulating factor receptor involved in induction of proliferation and maturation. *Mol Cell Biol*. 1993;13(12):7774-7781.
11. Fukunaga R, Ishizaka-Ikeda E, Nagata S. Growth and differentiation signals mediated by different regions in the cytoplasmic domain of granulocyte colony-stimulating factor receptor. *Cell*. 1993;74(6):1079-1087.
12. Germeshausen M, Skokowa J, Ballmaier M, Zeidler C, Welte K. G-CSF receptor mutations in patients with congenital neutropenia. *Curr Opin Hematol*. 2008;15(4):332-337.
13. Freedman MH, Bonilla MA, Fier C, et al. Myelodysplasia syndrome and acute myeloid leukemia in patients with congenital neutropenia receiving G-CSF therapy. *Blood*. 2000;96(2):429-436.
14. Link DC, Kunter G, Kasai Y, et al. Distinct patterns of mutations occurring in de novo AML versus AML arising in the setting of severe congenital neutropenia. *Blood*. 2007;110(5):1648-1655.
15. Clinical picture: AML (Acute Myeloid Leukemia). Trial: HOVON 42 A AML / SAKK.
16. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1617-1628.
17. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311.
18. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.
19. Hermans MH, van de Geijn GJ, Antonissen C, et al. Signaling mechanisms coupled to tyrosines in the granulocyte colony-stimulating factor receptor orchestrate G-CSF-induced expansion of myeloid progenitor cells. *Blood*. 2003;101(7):2584-2590.
20. Palande K, Roovers O, Gits J, et al. Peroxiredoxin-controlled G-CSF signalling at the endoplasmic reticulum-early endosome interface. *J Cell Sci*. 2011;124(Pt 21):3695-3705.
21. Carbuccia N, Trouplin V, Gelsi-Boyer V, et al. Mutual exclusion of ASXL1 and NPM1 mutations in a series of acute myeloid leukemias. *Leukemia*. 2010;24(2):469-473.

22. Taketani T, Taki T, Takita J, et al. AML1/RUNX1 mutations are infrequent, but related to AML-M0, acquired trisomy 21, and leukemic transformation in pediatric hematologic malignancies. *Genes Chromosomes Cancer*. 2003;38(1):1-7.

23. Morin RD, Mendez-Lago M, Mungall AJ, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*. 2011.

24. Pasqualucci L, Dominguez-Sola D, Chiarenza A, et al. Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature*. 2011;471(7337):189-195.

25. Score J, Hidalgo-Curtis C, Jones AV, et al. Inactivation of polycomb repressive complex 2 components in myeloproliferative and myelodysplastic/myeloproliferative neoplasms. *Blood*. 2011.

26. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.

27. Layton JE, Hall NE, Connell F, Venhorst J, Treutlein HR. Identification of ligand-binding site III on the immunoglobulin-like domain of the granulocyte colony-stimulating factor receptor. *J Biol Chem*. 2001;276(39):36779-36787.

28. Ohshiro T, Yagami T, Zhang C, Matsuzaki F. Role of cortical tumour-suppressor proteins in asymmetric division of Drosophila neuroblast. *Nature*. 2000;408(6812):593-596.

29. Gateff E. Malignant neoplasms of genetic origin in Drosophila melanogaster. *Science*. 1978;200(4349):1448-1459.

30. Peng CY, Manning L, Albertson R, Doe CQ. The tumour-suppressor genes lgl and dlg regulate basal protein targeting in Drosophila neuroblasts. *Nature*. 2000;408(6812):596-600.

31. Ito T, Kwon HY, Zimdahl B, et al. Regulation of myeloid leukaemia by the cell-fate determinant Musashi. *Nature*. 2010;466(7307):765-768.

32. Kharas MG, Lengner CJ, Al-Shahrour F, et al. Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat Med*. 2010;16(8):903-908.

33. Benz C, Mulindwa J, Ouna B, Clayton C. The Trypanosoma brucei zinc finger protein ZC3H18 is involved in differentiation. *Mol Biochem Parasitol*. 2011;177(2):148-151.

34. Mizuki M, Fenski R, Halfter H, et al. Flt3 mutations from patients with acute myeloid leukemia induce transformation of 32D cells mediated by the Ras and STAT5 pathways. *Blood*. 2000;96(12):3907-3914.

35. Shuai K, Halpern J, ten Hoeve J, Rao X, Sawyers CL. Constitutive activation of STAT5 by the BCR-ABL oncogene in chronic myelogenous leukemia. *Oncogene*. 1996;13(2):247-254.

36. Forbes LV, Gale RE, Pizzey A, Pouwels K, Nathwani A, Linch DC. An activating mutation in the transmembrane domain of the granulocyte colony-stimulating factor receptor in patients with acute myeloid leukemia. *Oncogene*. 2002;21(39):5981-5989.

37. Plo I, Zhang Y, Le Couedic JP, et al. An activating mutation in the CSF3R gene induces a hereditary chronic neutrophilia. *J Exp Med*. 2009;206(8):1701-1707.

38. McLemore ML, Poursine-Laurent J, Link DC. Increased granulocyte colony-stimulating factor responsiveness but normal resting granulopoiesis in mice carrying a targeted granulocyte colony-stimulating factor receptor mutation derived from a patient with severe congenital neutropenia. *J Clin Invest*. 1998;102(3):483-492.

39. Ward AC, Smith L, de Koning JP, van Aesch Y, Touw IP. Multiple signals mediate proliferation, differentiation, and survival from the granulocyte colony-stimulating factor receptor in myeloid 32D cells. *J Biol Chem*. 1999;274(21):14956-14962.

40. Pratcorona M, Abbas S, Sanders M, et al. Acquired mutations in ASXL1 in acute myeloid leukemia: prevalence and prognostic value. *Haematologica*. 2011.

41. Ernst T, Chase AJ, Score J, et al. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat Genet*. 2010;42(8):722-726.

42. Nikoloski G, Langemeijer SM, Kuiper RP, et al. Somatic mutations of the histone methyltransferase gene EZH2 in myelodysplastic syndromes. *Nat Genet*. 2010;42(8):665-667.

43. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011;478(7367):64-69.

CHAPTER

A single oncogenic enhancer rearrangement causes concomitant *EVI1* and *GATA2* deregulation in leukemia

Stefan Gröschel[1,2,*], Mathijs A. Sanders[1,*], Remco Hoogenboezem[1], Elzo de Wit[3], Britta A.M. Bouwman[3], Claudia Erpelinck[1], Vincent H.J. van der Velden[4], Marije Havermans[1], Roberto Avellino[1], Kirsten van Lom[1], Elwin J. Rombouts[1], Mark van Duin[1], Konstanze Döhner[2], H. Berna Beverloo[5,6], James E. Bradner[7,8], Hartmut Döhner[2], Bob Löwenberg[1], Peter J.M. Valk[1], Eric M.J. Bindels[1], Wouter de Laat[3], and Ruud Delwel[1]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands
[2] University Hospital Ulm, Department of Internal Medicine III, Ulm, Germany
[3] Hubrecht Institute-KNAW and University Medical Center Utrecht, Utrecht, The Netherlands
[4] Erasmus University Medical Center, Department of Immunology, Rotterdam, The Netherlands
[5] Erasmus University Medical Center, Department of Clinical Genetics, Rotterdam, The Netherlands
[6] Dutch Working Group on Hemato-Oncologic Genome Diagnostics
[7] Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, USA
[8] Harvard Medical School, Department of Medicine, Boston, USA

*These authors contributed equally to this work

Cell 2014 April 10;157(2):369-81

ABSTRACT

Chromosomal rearrangements without gene-fusions have been implicated in leukemogenesis by causing deregulation of proto-oncogenes via relocation of cryptic regulatory DNA elements. AML with inv(3)/t(3;3) is associated with aberrant expression of the stem-cell regulator *EVI1*. Applying functional genomics and genome-engineering, we demonstrate that both 3q-rearrangements reposition a distal *GATA2* enhancer to ectopically activate *EVI1* and simultaneously confer *GATA2* functional haploinsufficiency, previously identified as the cause of sporadic familial AML/MDS and MonoMac/Emberger syndromes. Genomic excision of the ectopic enhancer restored *EVI1* silencing and led to growth inhibition and differentiation of AML cells, which could be replicated by pharmacologic BET-inhibition. Our data show that structural rearrangements involving single chromosomal repositioning of enhancers can cause deregulation of two unrelated distal genes, with cancer as the outcome.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter7/

INTRODUCTION

Chromosomal inversions and translocations play a central role in the pathogenesis of almost all types of cancers, frequently by formation of oncogenic fusion genes via rearrangement of coding sequences of the involved partner genes.[1-3] Mechanisms of transformation remain largely unknown in malignancies arising from chromosomal inversions/translocations that do not cause fusion products, although it is thought that destabilization of cryptic regulatory elements affects genes in the vicinity of the structural rearrangement, as has been shown in Burkitt's[4] or follicular lymphoma.[5,6]

In the World Health Organization (WHO) category of myeloid malignancies with inv(3)(q21q26.2) or t(3;3)(q21;q26.2), deregulation of the proto-oncogene *EVI1* (also termed *MECOM* or *PRDM3*) at 3q26.2 is speculated to occur via juxtaposition of a cryptic enhancer of the housekeeping gene *RPN1* from 3q21.[7] However, this hypothesis has not been experimentally validated and the molecular basis of this prognostically unfavorable subtype of malignancies remains obscured. *EVI1* expression and function is indispensable for proper regulation of the hematopoietic stem cell compartment and genomic integrity.[8-10] The gene was originally described as a hotspot for proviral integration in retrovirally induced murine myeloid leukemias[11], and also represents an important insertional mutagenesis site in humans following gene therapy for X-linked granulomatous disease.[12]

We tested the hypothesis that rearrangements causing the transcriptional activation of *EVI1* involve the reallocation of an enhancer element to the ectopic 3q26.2/*EVI1* target site, which may possibly coincide with a loss of enhancer activity at its endogenous location. We applied an integrated functional genomics and genome-engineering approach to identify a distal enhancer of the *GATA2* gene that, upon chromosomal 3q-rearrangements, ectopically activates *EVI1* expression. Simultaneously, the removal of this enhancer from its natural genomic context causes functional *GATA2* haploinsufficiency, i.e. reduced *GATA2* expression only from the remaining normal allele.

RESULTS

An 18 kb non-coding region near *RPN1* commonly translocates to *EVI1* in inv(3)/t(3;3) disease

We performed targeted next generation sequencing (NGS) of the long arm of chromosome 3 (3q-seq) using genomic DNA isolated from 41 samples with confirmed *EVI1* overexpression (*EVI1*$^{+}$) and harboring an inv(3)(q21q26.2) or a t(3;3)(q21;q26.2) [inv(3)/t(3;3)]. The samples included 38 primary bone marrow samples from patients, i.e. AML (n=33), CML-BC (n=2), and MDS cases (n=3), as well as three cell lines (MUTZ-3, MOLM-1, and UCSD-AML1) (Table S1). Chromosomal breakpoint positions and novel junction sequences of each case were determined by a breakpoint detection

algorithm in conjunction with a de novo assembly tool and validated by Sanger sequencing. Characteristic breakpoint patterns emerged at both 3q21 and 3q26.2 breakpoint cluster regions (Figure 1A). At the 3q26.2 site, samples harboring an inv(3) exclusively exhibited breakpoints in the last intron or downstream of *EVI1*. Breakpoints in t(3;3) cases distinctly mapped upstream of *EVI1,* i.e. within the gene locus of the longer splice variant that includes *MDS1-EVI1* (Figure 1A). At the 3q21 site, breakpoints occurred in a 130 kb region between *GATA2* (centromeric) and *RPN1* (telomeric). A minimal 18 kb non-coding region 3' of *RPN1* demarcated by chromosomal breakpoints was identified as a commonly translocated segment (hereafter referred to as CTS) (Figure 1B), which in all cases underwent transpositioning to the vicinity of *EVI1* due to the inv(3)/t(3;3) rearrangement. This converging tell-tale pattern of 3q21 breakpoints toward an unaffected 18 kb genomic segment led us to predict the presence of potent regulatory elements within the CTS, essential for aberrant activation of *EVI1* upon rearrangement.

The *EVI1* promoter and the 18 kb CTS physically interact

A hallmark of distal enhancer elements is their engagement in chromatin loops physically contacting with promoters to induce transcription factor assembly and polymerase II recruitment.[13-16] To test whether the CTS harbored elements physically interacting with the *EVI1* promoter, we performed high-resolution chromosome conformation capture sequencing (4C-seq) experiments.[17] Using viewpoints placed on the *EVI1* promoter in viable inv(3)/t(3;3) AML samples and cell lines, we identified a genomic segment of approximately 9 kb size within the 18 kb CTS contacting with the *EVI1* promoter (Figures 1C and 1D). Other contact regions located centromeric of this 9 kb contact hotspot in closer distance to the *EVI1* promoter after the rearrangement, as observed in individual samples with different breakpoint positions, were considered less likely enhancer candidates. These regions were non-overlapping across different samples and thus represented less specific contacts, which became more evident after integrative analysis of all 3q-rearranged AML samples (Figure 1D). Reciprocal 4C-seq experiments with the putative 9kb region as viewpoint showed that the interaction with the *EVI1* promoter area was also evident (Figure S1). As expected, no substantial chromatin interactions with the distant *EVI1* promoter could be detected in 4C-seq experiments with non-3q-rearranged control (U937), suggesting an inv(3)/t(3;3) disease-specific feature (Figures 1C and S1).

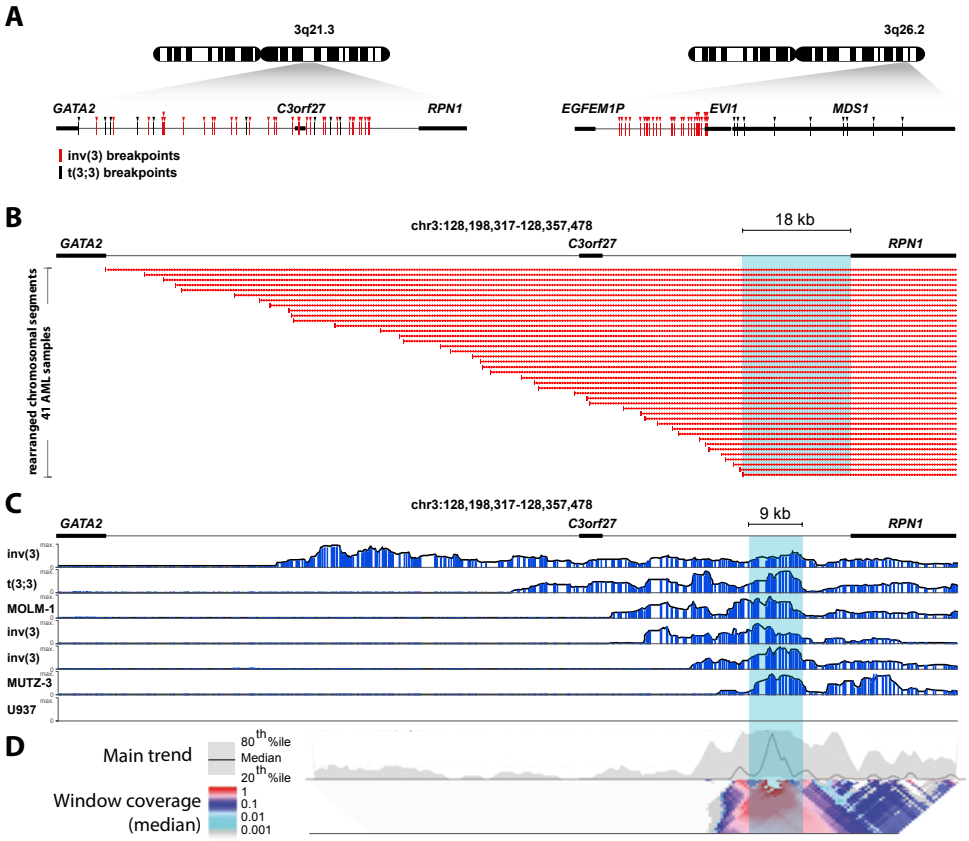


Figure 1. Chromosomal breakpoint architecture in inv(3)/t(3;3) AML and local 3q21 chromatin interaction profiles of the *EVI1* promoter. (A) Mapping of chromosomal breakpoints (arrowheads) by targeted 3q-capture NGS revealing two characteristic breakpoint patterns at 3q21 (left panel) and 3q26.2 (right panel). (B) The 3q21 breakpoint cluster and rearranged chromosomal segments of individual AML samples are represented by red arrowed lines, plotted by distance to the *RPN1* gene locus. A breakpoint-free commonly translocated segment (CTS) of 18 kb size is indicated (blue box). (C) The local chromatin interaction profile of the *EVI1* promoter region with the 3q21 breakpoint cluster was determined by 4C-seq in representative inv(3)/t(3;3) cases. The 4C signal is measured by the calculation of a sliding window average (running mean) of the normalized read counts (window size is 21 fragment ends). The vertical axis is scaled to the maximum 4C signal per sample. An overlapping contact hotspot of 9 kb size within the CTS in 3q-rearranged cases is highlighted as a blue box. The non-3q-rearranged cell line U937 was used as control. (D) Integrated local contact profile analysis of all 3q-rearranged samples. In the top panel (main trend), the contact intensity (black line) is calculated by using a running median analysis of normalized read counts with a 5 kb sliding window. The 20th and 80th percentile are visualized as a grey trend graph. In the bottom panel, contact intensities are computed using linearly increasing sliding windows (scaled 2-50 kb) and are displayed as a color-coded heatmap of positive 4C signals (maximum of interaction set to 1). Local color changes are log-scaled to indicate changes of statistical enrichment of captured sequences, corresponding to the enhancer-promoter interaction.

A p300-bound genomic element in the 18kb CTS is essential for *EVI1* activation

In order to identify a more defined, targetable key enhancer element within the 9 kb *EVI1*-promoter-contact part of the 18 kb CTS, we integrated data from 4C-seq with ChIP-seq data obtained from inv(3) cell lines MOLM-1 and MUTZ-3 (Figures 2A and S2). Prominent deposition of H3K27ac, H3K4me3, and H3K4me1 was observed within the 18 kb CTS, as well as strong binding of p300 to two regions of approximately 1 kb size in MOLM-1 (Figure 2A).

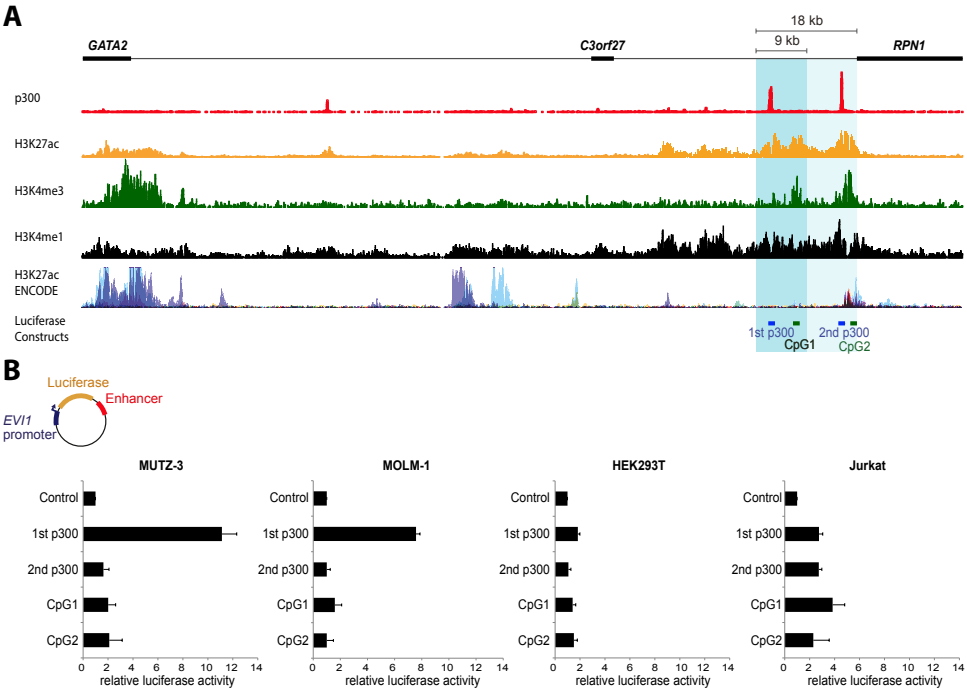
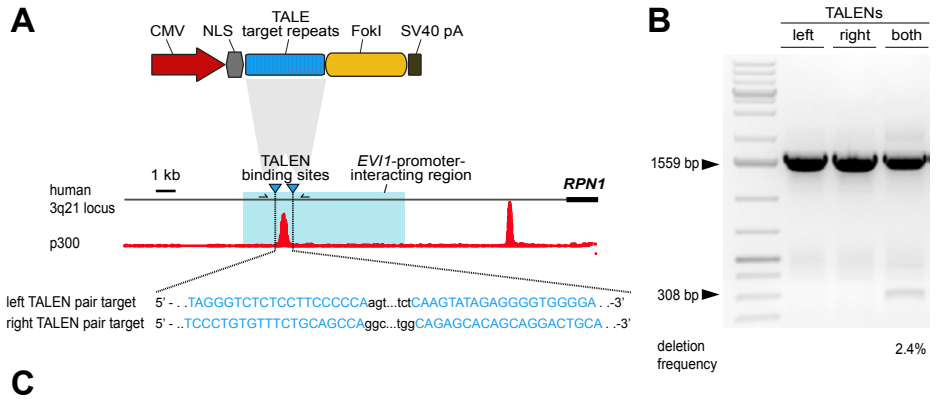


Figure 2. Combined ChIP-seq and 4C-seq discloses putative enhancer elements in the *EVI1* promoter interacting rearrangement site. (A) Binding of p300 (red track) in the 9 kb *EVI1*-interaction domain of the 18 kb CTS (divided blue box) is detectable in inv(3)/t(3;3) samples. ChIP-seq profiles of p300 and active chromatin marks H3K27ac, H3K4me3, and H3K4me1 of the inv(3) cell line MOLM-1 indicate an inv(3) cell-type specific enrichment in the CTS, not found in non-3q rearranged cell lines of various tissue origin (ENCODE: GM12878, H1-hESC, HeLa-S3, HepG2, HSMM, HUVEC, K562, NHEK, NHLF). Based on ChIP-seq enhancer profiles and CpG island prediction, two candidate enhancer regions were selected both inside (denoted 1st p300 and CpG1; blue and green bars) and outside (denoted 2nd p300 and CpG2; blue and green bars) of the 9 kb *EVI1*-interaction domain of the CTS, respectively, for subsequent reporter assays. (B) Selected candidate elements were cloned into *EVI1*-promoter luciferase reporter constructs and transfected into MUTZ-3, MOLM-1, HEK293T, or Jurkat cells. After 48 h, cells were harvested and luciferase assays were performed. Relative luciferase induction is plotted as fold change compared to enhancer-empty control vector (mean ± s.e.m.).

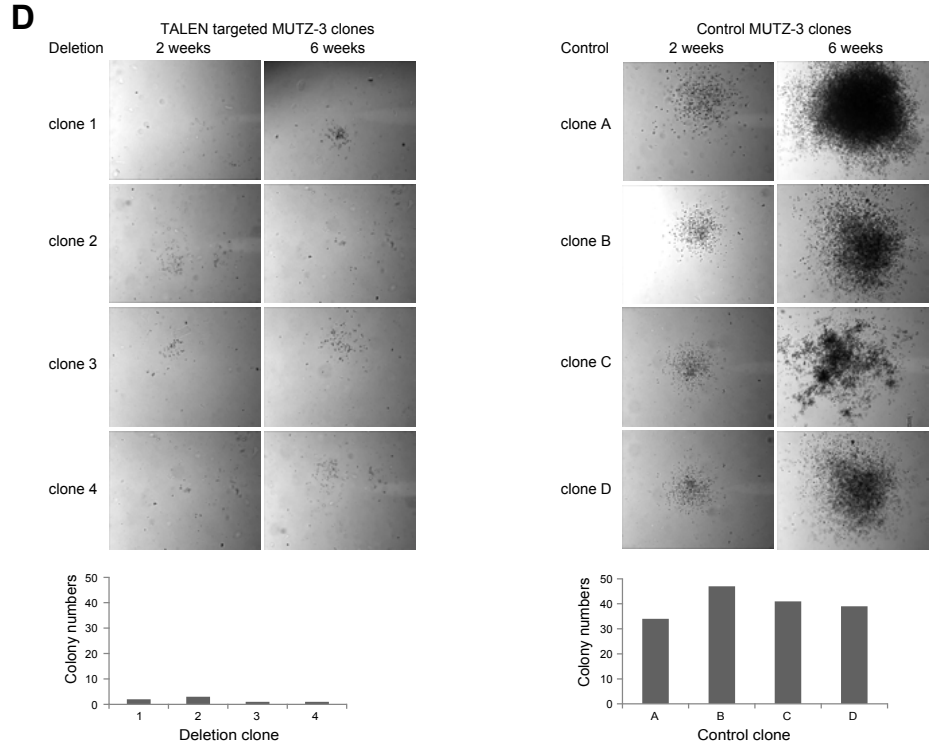
In MUTZ-3 cells only one p300-interacting region was identified (Figure S2), which is the p300 peak located in the 9kb *EVI1*-promoter-contact part of the CTS as determined by 4C-seq. Comparison with ENCODE ChIP-seq data of various non-myeloid cell lines and transcription factor motif analysis pointed to a 1 kb myeloid hematopoiesis-specific enhancer (Figures 2A and S2). This p300 binding site was chosen as the most likely candidate enhancer element responsible for ectopic *EVI1* activation after the rearrangement event. We placed the two candidate enhancers (1st and 2nd p300) into an *EVI1*-promoter luciferase reporter construct to study their potential enhancer activity (Figure 2B). A strong induction of reporter gene activity could be observed specifically in inv(3) myeloid cell lines MUTZ-3 and MOLM-1 using the first candidate enhancer element, whereas the second candidate enhancer element (2nd p300 peak) located within the 18 kb CTS, but outside of the 9kb *EVI1*-promoter-contact region, showed no enhancer activity. No activity was found for two distinct CTCF-interacting CpG islands co-localizing to the CTS. Moreover, *EVI1* promoter reporter assays demonstrated no substantial enhancing effect of the candidate 1 kb enhancer in non-myeloid HEK293T or Jurkat cells, pointing to a myeloid-specific transcription factor repertoire required for successful enhancer-*EVI1*-promoter engagement.

Genome-editing of the translocated p300-interaction domain leads to *EVI1* silencing and growth inhibition of inv(3) AML cells

We next undertook a TALE nuclease genome-editing approach to target the ectopic *EVI1* enhancer locus in the MUTZ-3 cell line and to examine whether *EVI1* transcriptional activity in inv(3) AML cells is dependent on the presence of the rearranged candidate enhancer (1st p300 peak). TALE nucleases were assembled as previously published[18], and targeting of the minimal ectopic enhancer site was performed in a 2x2 design (details in Experimental procedures section), directing TALEN heterodimers to enhancer-flanking recognition sequences to induce a segmental deletion by double-strand breaks (DSB) (Figure 3A). Mutation-specific primers allowed for allelic detection of the deletion event (Figure 3B) and for screening of clones using an informative SNV in the candidate enhancer locus of MUTZ-3 (Figures 3B, S3A, and S3B). Overall targeting efficiency was 1% with 4/384 single-cell derived clones harboring a monoallelic enhancer deletion on the inv(3) allele (Figure S3C). Enhancer-targeting effectively attenuated *EVI1* mRNA expression in deletion clones as compared with non-targeted MUTZ-3 control clones taken along in the same targeting process (Figure 3C). *RPN1* and *GATA2* mRNA expression remained unchanged in inv(3)-targeted clones. Notably, all four TALEN-modified MUTZ-3 clones showed severely impaired colony-forming and replating capacity compared to non-targeted clones (Figure 3D).


A
CMV
NLS
TALE target repeats
FokI
SV40 pA
TALEN binding sites
EVI1-promoter-interacting region
1 kb
human 3q21 locus
RPN1
p300
left TALEN pair target 5' - . .TAGGGTCTCTCCTTCCCCCAagt...tctCAAGTATAGAGGGGTGGGGA . .-3'
right TALEN pair target 5' - ..TCCCTGTGTTTCTGCAGCCAggc...tggCAGAGCACAGCAGGACTGCA . .-3'
B
TALENs
left
right
both
1559 bp
308 bp
deletion frequency
2.4%


C
EVI1
GATA2
RPN1
normalized mRNA levels
Deletion clone
Control clone

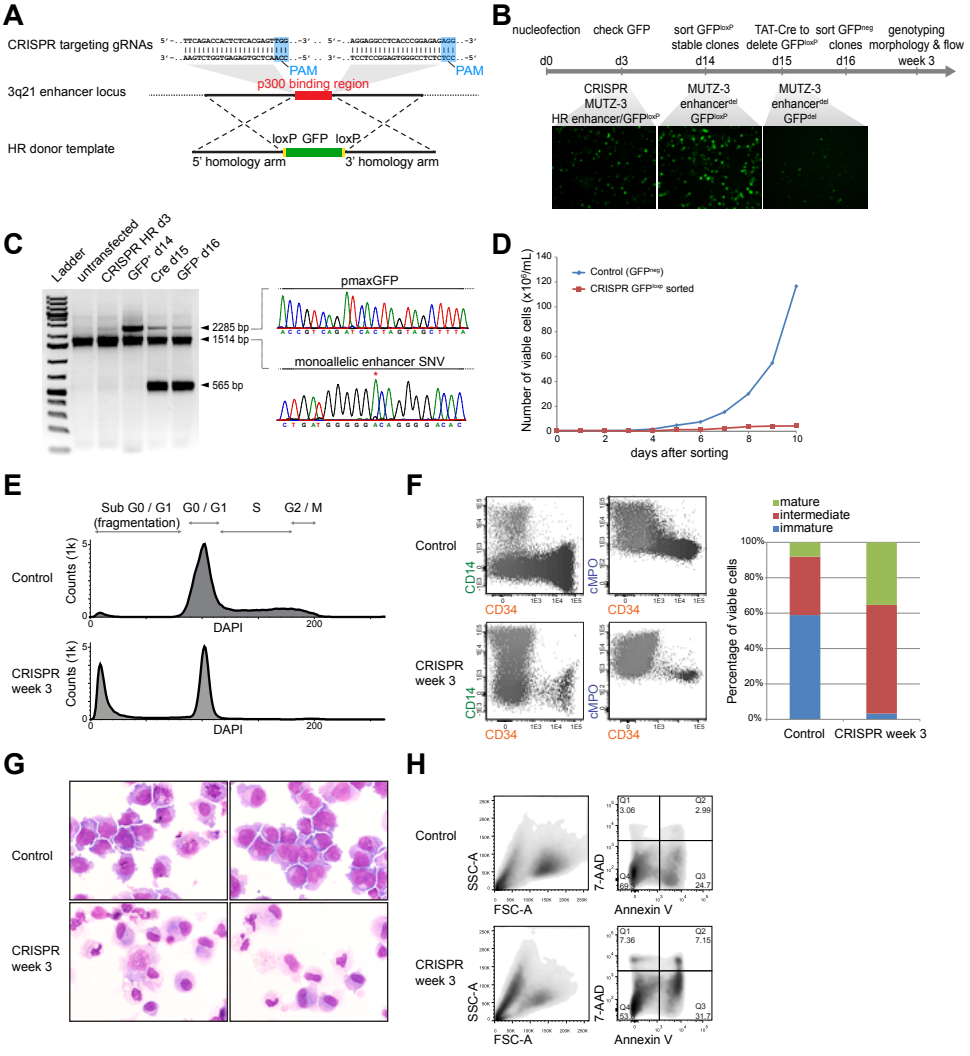

D
TALEN targeted MUTZ-3 clones
Deletion
2 weeks
6 weeks
clone 1
clone 2
clone 3
clone 4
Colony numbers
Deletion clone
Control MUTZ-3 clones
Control
clone A
clone B
clone C
clone D
Control clone

Deletion of the ectopic *EVI1* enhancer releases the maturation block of inv(3) AML cells

Genome-editing in MUTZ-3 AML cells using TALENs did not allow for high yields of viable cells lacking the enhancer because of the low targeting efficiency of this genetic tool in the inv(3) cell line model followed by a long and indirect selection process of growth-impaired deletion clones. To better characterize the cellular phenotype and fate after enhancer deletion, we designed an alternative targeting approach using the CRISPR/Cas9 genome-editing system with a site-specific homology repair (HR) donor for direct labeling and tracking of successfully targeted cells (Figures 4A and 4B). In brief, enhancer deletions were induced by two short guide-RNAs (gRNA) directing hSpCas9 to two enhancer-flanking recognition sequences for DSB formation and HR-mediated repair of the induced segmental deletion by means of a co-transfected HR donor construct containing a conditional (loxP-) GFP selection cassette directed against the enhancer (details in Experimental procedures section). The pCMV-GFP cassette was subsequently removed by using exogenous cell-permeant Cre recombinase (TAT-Cre). This approach enabled us to isolate sufficient cell numbers for phenotypic characterization. Deletion events and transcriptional changes were confirmed by Sanger sequencing and qPCR (Figures 4C and S3D). Compared to untargeted control MUTZ-3 cells, targeted cells exhibited a markedly reduced proliferative rate as assessed by viable cell count using trypan blue staining (Figure 4D). Cell cycle analysis showed depletion of S phase and G2/M phase combined with higher rates of cell death (sub-G0/G1 peak) and a stationary G0/G1 cycle arrest (Figure 4E). Remarkably, flow cytometric immunophenotyping of engineered cells using a panel of informative myeloid differentiation markers (see Supplemental Experimental Procedures for detailed list) according to published guidelines[19] revealed a substantial skew toward a more mature, myelomonocytic stage as per cMPO and CD14 expression levels three weeks after genome-editing (Figure 4F). Cytologic evaluation of enhancer-targeted MUTZ-3 cells in week 3 after genomic modification confirmed morphologic changes from the predominantly immature, myelomonocytic appearance of untargeted cells toward a more differentiated, monocyte/macrophage-like shape (Figure 4G). This also translated into a higher apoptotic rate of CRISPR-targeted MUTZ-3 cells three weeks after enhancer deletion (Figure 4H).

Figure 3. TALEN-targeted candidate enhancer-deletion clones exhibit severely reduced *EVI1* mRNA levels and replating capacity. (A) Schematic of the targeting construct and target sites flanking the previously identified p300-binding candidate enhancer (red ChIP-seq track) in the 9 kb *EVI1*-interaction domain of the CTS (blue box). Four TALENs were designed for pairwise heterodimeric binding to indicated target sequences (blue) and cleavage at the 16-18 bp intervening linker sequence (black). Arrows indicate primer locations for PCR analysis. (B) Representative gel image demonstrating efficient induction of segmental deletions only in the presence of two TALEN pairs (MUTZ-3 cell line; 2.4% targeting efficiency as per gel quantification 48 h after transfection). (C) *EVI1*, *GATA2*, and *RPN1* mRNA expression analysis by qPCR of genome-edited MUTZ-3 mutants after TALEN-targeting of the candidate enhancer on inv(3). (D) Comparison of colony formation of targeted and unmodified clones. Colony numbers and sizes were determined after two and six weeks after replating in methylcellulose.

Off-target mutagenesis at alternative in silico predicted sites was excluded by Sanger sequencing (Figure S3E). The phenotype observed upon enhancer deletion by genome-editing tools was highly comparable to what was found with small hairpin RNA (shRNA)-mediated *EVI1* knockdown in the MUTZ-3 cell line (Figures S4A-S4H), emphasizing that MUTZ-3 cells are addicted to *EVI1* and blocked in their differentiation. We did not observe outgrowth of biallelic enhancer deletion or monoallelic mutants of the non-rearranged chromosome 3 allele, hinting toward negative selection of these mutants upon disruption of the enhancer at its natural genomic location.



The candidate enhancer translocated to *EVI1* is an original upstream enhancer of *GATA2*

The most likely candidate for off-target mutagenesis at the original enhancer-associated domain on the normal chromosome 3 allele appeared to be *RPN1* due to its immediate proximity to the candidate enhancer. Concordantly, *RPN1* has therefore generally been the assumed origin of ectopic *EVI1* regulatory elements, since it is located in the proximity of the chromosome 3 breakpoint cluster.[7,20] Thus, disturbance of the housekeeping function of *RPN1* on the remaining normal allele in our TALENs experiment would most likely be deleterious. However, analysis of our 4C-seq profiling data instead revealed substantial interactions between the candidate enhancer and the promoter of *GATA2* rather than with the promoter of *RPN1* (Figures 5A, 5B, and S5A). *GATA2* is a crucial hematopoietic stem cell regulator, located on the contralateral side of the 3q21 breakpoint cluster. This suggests that the candidate enhancer is an original upstream regulatory element for *GATA2*, rather than *RPN1*. Hi-C data confirm that the candidate enhancer is together with the *GATA2* locus in a topological domain, physically segregated from the more proximal *RPN1* promoter (Figure S5B).[21]

Consequently, we first aimed to examine the effect of loss of the candidate enhancer in a human cell line without 3q-rearrangements and the functional impact on either *RPN1* or *GATA2* expression. We generated custom CRISPR/Cas9 nucleases against the candidate enhancer locus in the *GATA2*-expressing erythroleukemia cell line K562 (Figure 6A). We effectively deleted the candidate enhancer in K562 cells and observed profoundly reduced levels (10.8-fold) of *GATA2* mRNA in targeted K562 pools (Figure 6B), whereas *RPN1* expression levels remained unchanged. Luciferase *GATA2*-promoter reporter studies confirmed strong *GATA2*-specific enhancer activity of the candidate locus in a myeloid context (Figure 6C). Thus, the candidate ectopic enhancer, which upon translocation is repositioned to the *EVI1* locus, is a strong enhancer of *GATA2* in its original chromosomal context.

Figure 4. Genomic enhancer excision induces proliferative and differentiation changes in inv(3) AML cells. (A) Schematic representation of the CRISPR/Cas9 licensing gRNAs with protospacer-adjacent motifs (PAM) highlighted in blue, the target locus, and the donor construct for site-directed homology repair using a conditional, floxed pCMV-GFP selection cassette. (B) Timeline of genomic targeting of MUTZ-3 AML cells. (C) Detection of deletion events by genomic PCR of sequential cell fractions. Representative Sanger sequencing tracks of purified PCR amplicons of the GFP-insertion band (2.3 kb) and a remaining lower-running, normal allele band of 1.5 kb size are shown (from GFP^+ fraction of day 14), revealing a monoallelic deletion indicated by a loss of heterozygosity of the SNV present in the targeted enhancer locus (red asterisk). (D) Proliferation of untargeted control and targeted cells was measured by counting of viable cells using trypan blue. (E) Cell cycle analysis of control and genome-edited MUTZ-3 cells harvested after three weeks of selection. (F) Immunophenotyping of control and enhancer-targeted MUTZ-3 cells. The left panel includes two dot plots per sample (CD34/CD14 and CD34/cMPO) that show the myelomonocytic maturation. The right panel shows the distribution of the various maturation stages, simplified in three stages: immature ($CD34^+/CD14^-$) = blast cell; intermediate ($CD34^-/CD14^-$) = promonocyte; mature ($CD34^-/CD14^+$) = monocyte. (G) Representative images of May-Grünwald-Giemsa staining of control and enhancer-targeted MUTZ-3 cells (100x magnification). (H) Assessment of apoptosis in control (top panels) and enhancer-targeted MUTZ-3 cells (bottom panels). Representative flow cytometry plots for Annexin V and 7-AAD staining with percentages for each gate are shown.

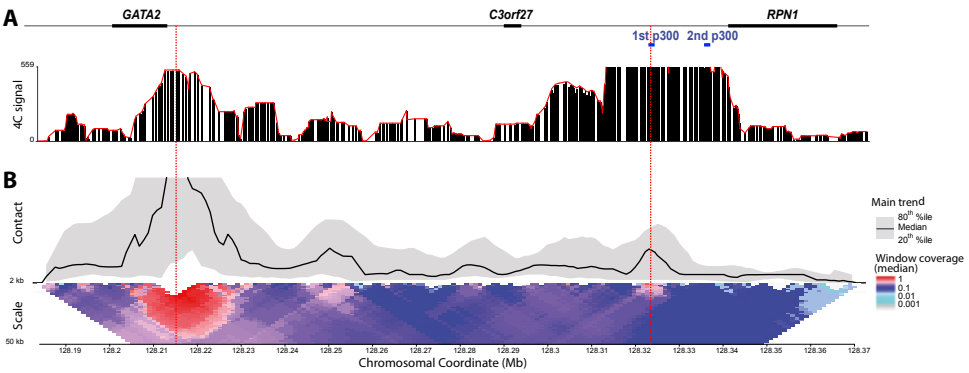


Figure 5. The non-rearranged candidate enhancer is part of the *GATA2* enhancer complex. (A) Representative 4C data (n=7 biological replicates) showing the local contact profile using a window of 21 fragment ends with the 1st p300 peak site as viewpoint (red dashed line) (B) Integrative 4C analysis using a viewpoint from the *GATA2* promoter region (n=7 biological replicates). In the top panel (main trend), the contact intensity (black line) is calculated by using a running median analysis of normalized read counts with a 5 kb sliding window. The 20th and 80th percentile are visualized as a grey trend graph. In the bottom panel, contact intensities are computed using linearly increasing sliding windows (scaled 2-50 kb) and displayed as a color-coded heatmap of positive 4C signal (maximum of interaction set to 1). Local color changes are log-scaled to indicate changes of statistical enrichment of captured sequences, corresponding to the enhancer-promoter interaction (red dashed lines).

Rearrangement of the *GATA2* enhancer to *EVI1* causes functional haploinsufficiency of *GATA2*

To study the effects of the enhancer translocation on *GATA2* expression in inv(3)/t(3;3) AML samples, we analyzed allele frequencies of informative SNPs in the *GATA2* locus by combining 3q-seq and RNA-seq data. This integrative analysis revealed a monoallelic expression pattern of *GATA2* in all 36 inv(3)/t(3;3) cases studied (Figures 6D and S6). Non-3q-rearranged AML patient samples and cell lines, as well as variant t(3q26) AML cases with breakpoints outside of the 3q21 cluster region [e.g., inv(3)(q21q25); t(3;7)(q26;p15)] displayed a normal, biallelic *GATA2* expression pattern (not shown). To ascertain monoallelic *GATA2* expression originating from the normal chromosome 3 allele, we performed an allele-specific chromosome conformation capture sequencing approach (Experimental procedures for details), in which captured informative SNPs of the *GATA2* locus can only be amplified by allele-specific primers on the non-rearranged, linear chromosome 3 template. Results were validated by long-range, breakpoint-spanning PCR and Sanger sequencing. By integration of results from these NGS platforms (3q-seq, RNA-seq, and allele-specific 4C; Figure 6D), we found monoallelic *GATA2* expression as a consequence of *GATA2* inactivation on the rearranged allele in cases harboring inv(3) or t(3;3). Notably, *GATA2* expression levels in primary inv(3)/t(3;3) AML cases and cell lines (n=69) were found to be significantly reduced as compared to controls (213 AML patients) (Figure 6E). Thus, our data indicate that the inversion/translocation event in inv(3)/t(3;3) malignancies reorganizes an originally upstream regulatory element of the *GATA2* domain, causing reduced and monoallelic expression of *GATA2*.

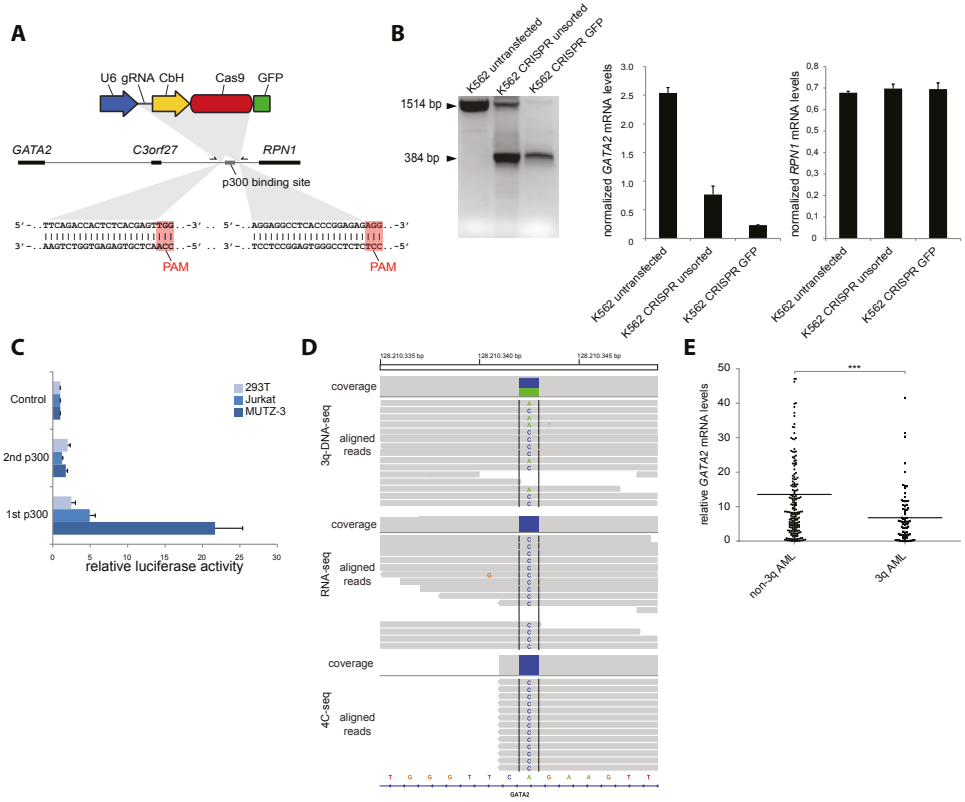


Figure 6. Loss of the *GATA2* candidate enhancer leads to functional haploinsufficiency of the affected *GATA2* allele. (A) Schematic of the CRISPR nuclease design for candidate enhancer targeting. Arrows indicate primer locations for PCR analysis. For each construct, the protospacer sequence and the Cas9-specific proximal-adjacent motif (PAM; magenta highlight) are indicated. (B) Upon transfection of the candidate enhancer-flanking CRISPR constructs, K562 cells were analyzed by deletion-specific PCR (gel image). Unsorted cells represented pools of CRISPR-targeted and non-targeted cells. GFP-sorted and isolated deletion clones harbored predominantly biallelic deletion mutants. *GATA2* and *RPN1* mRNA expression was analyzed by qPCR (right panel). (C) The p300-binding core enhancer region and an adjacent control region (2nd p300 peak region) were cloned into a *GATA2*-promoter luciferase reporter construct, and luciferase activity was measured 48 h after transfection of indicated cell lines. *GATA2*$^{+}$ MUTZ-3 cells, as well as *GATA2* non-myeloid HEK293T and Jurkat cells were assayed. (D) Integrated analysis of 3q-DNA-seq, RNA-seq, and allele-specific 4C-seq data of a representative inv(3) AML case reveals monoallelic expression of *GATA2* mRNA from the intact chromosome 3q21 allele. (E) *GATA2* expression level analysis by qPCR in inv(3)/t(3;3) AML (n=69) and unselected, non-3q-rearranged AMLs (n=245; Mann-Whitney-U test, p=0.002).

The 18 kb CTS and p300-interaction domain are part of a translocation-derived super-enhancer

We have shown that targeting of the candidate enhancer site in inv(3) AML cells by genome-editing techniques is feasible, based on the premises that the enhancer element interacts with the *EVI1* promoter, binds the transcriptional co-activator p300, and is embedded in a region of

open, potentially regulatory chromatin, and thereby also accessible for endonucleases. However, ChIP-seq data obtained from the inv(3) cell line MOLM-1 manifested a large segment of H3K27ac deposition that extends beyond the entire 18 kb CTS and p300-interaction domain, covering a region of approximately 40 kb (Figures 2A and 7A). These exceptionally large enhancer domains with high levels of H3K27ac and the chromatin regulator BRD4 have recently been characterized as super-enhancers.[22,23] Using the bioinformatic analysis tool ROSE[22], the 40 kb H3K27ac-deposition region was identified as a super-enhancer, ranking second among 291 super-enhancers in the MOLM-1 genome (Figures 7A and 7B). RNA-seq analysis revealed the presence of intense read-through enhancer RNAs (eRNAs) spanning the entire super-enhancer area including the 18 kb CTS in MOLM-1 (Figure 7A). Of note, read-through transcription commenced precisely at the breakpoint positions, representing the fusion point of 3q21 with 3q26/*EVI1* segments. RNA-seq carried out in all available inv(3)/t(3;3) leukemia samples disclosed identical patterns of large read-through areas of eRNAs (Figure 7A). Consistently, BRD4-occupancy was found at the super-enhancer site in 3q-rearranged samples, particularly in the p300-interaction domain (Figure S7A). Non-3q-rearranged samples entirely lacked traces of transcriptional read-through (Figure 7A) and exhibited no H3K27ac deposition, or, if any at all, only in a confined 3-4 kb region immediately downstream of the *RPN1* gene, as shown by comparison with ENCODE ChIP-seq data of various non-3q-rearranged cell lines (Figures 2A and S2). Furthermore, combining 3q-capture DNA-seq with ChIP-seq data of MOLM-1 showed the presence of informative heterozygous SNPs in the putative 3q21 super-enhancer locus on genomic DNA level, whereas the chromatin after H3K27ac pull-down revealed a skew in the allelic ratio of these SNPs in the same locus (Figure 7C). These observations suggest the presence of an active, rearranged super-enhancer in inv(3)/t(3;3) leukemia samples, as was previously observed for *MYC*-rearrangements in multiple myeloma.[22]

BET-inhibition leads to *EVI1* silencing and growth arrest of inv(3)/t(3;3) AML cells

Our genome-editing results underline that *EVI1* is the key oncogenic driver in inv(3)/t(3;3) AML and vulnerable to interference with its ectopic enhancer. As reported previously, BET-bromodomain inhibition of super-enhancers represents a novel therapeutic avenue to target genes particularly regulated by super-enhancers.[22,23] The observation that the p300-binding ectopic *EVI1* enhancer is embedded in a large 3q21 super-enhancer complex (Figures 7A and 7B) prompted us to investigate whether *EVI1* transcription in inv(3)/t(3;3)-rearranged AMLs is sensitive to enhancer interference by treatment with BET-bromodomain inhibitors (JQ1). Exposure of MUTZ-3 and MOLM-1 cells, as well as primary inv(3)/t(3;3) AML samples to JQ1 profoundly inhibited proliferation with concentrations >50 nM (Figures 7D and S7C). *EVI1*-expressing K562 cells (no 3q-rearrangement), however, were not responsive to JQ1, as was previously shown.[24] *EVI1* mRNA levels in MUTZ-3 and MOLM-1 cells significantly decreased after JQ1 treatment contrary to K562 cells, in which BRD4 density at the enhancer locus is lower by comparison (Figures 7E, S7A, S7D, and S7E).

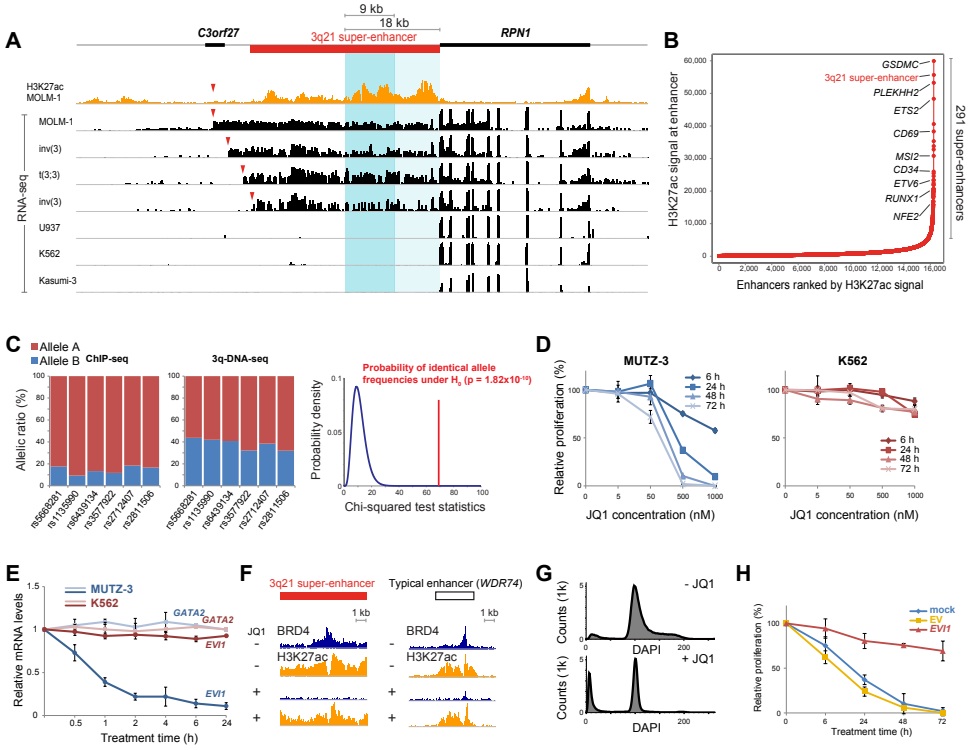


Figure 7. The ectopic enhancer is part of a 3q21 super-enhancer. (A) H3K27ac ChIP-seq (orange track) of the MOLM-1 cell line and RNA-seq (black tracks) of representative inv(3)/t(3;3) samples and non-3q-rearranged cell lines. Red arrowheads denote chromosomal breakpoint positions per sample. The super-enhancer region defined by the H3K27ac enrichment score is indicated as a red bar. (B) Ranking of super-enhancers identified in the MOLM-1 genome as per H3K27ac enrichment. (C) Allelic imbalance of the 3q21 super-enhancer region is determined by comparison of allelic ratios obtained from 3q-seq and H3K27ac ChIP-seq using informative, heterozygous SNPs (n=6) present in the super-enhancer domain. Allelic imbalances for SNPs are tested using a χ^2 goodness-of-fit test to identify regions exhibiting ChIP-seq allelic ratios significantly different from the genomic allelic ratios (H0 = allelic ratios identical between ChIP-seq and 3q-DNA-seq experiments). (D) JQ1 treatment is effective in 3q-rearranged AML cells (MUTZ-3) vs. non-3q-rearranged cells (K562). JQ1 sensitivity was measured by mitochondrial dehydrogenase (MTT assay) after 6, 24, 48, and 72 h of exposure with JQ1 (5, 50, 500, or 1,000 nM) or vehicle control (DMSO, 0.05%). (E) Analysis of *EVI1* and *GATA2* mRNA expression levels in MUTZ-3 and K562 by qPCR at different time points following JQ1 exposure (1,000 nM). (F) ChIP-seq binding profiles for BRD4 (blue) and H3K27ac (orange) at the 3q21 super-enhancer locus (left panel) or at the *WDR74* upstream enhancer after treatment of MOLM-1 cells with JQ1 (1,000 nM) or DMSO (0.05%) for six hours. (G) Cell cycle analysis of MUTZ-3 cells after treatment with DMSO (0.05%; upper panel) or JQ1 (1,000 nM; lower panel) for 24 h. (H) Transient *EVI1* rescue counteracts JQ1 antiproliferative effect. Cells were nucleofected either without DNA (mock), or with an empty GFP-expressing vector (EV) or an GFP-EVI1-expressing construct (*EVI1*), GFP-sorted after 24 h, and subsequently exposed to JQ1 (1,000 nM) for following viability measurements at indicated time points.

Furthermore, *GATA2* mRNA levels did not change upon JQ1 treatment, substantiating the notion that the ectopic super-enhancer/core p300 element is indeed a fusion result regulating *EVI1* on the rearranged allele rather than *GATA2* on the remaining normal allele. BRD4 load as well as

read-through transcription at the ectopic *EVI1* super-enhancer site were substantially decreased after JQ1 treatment (Figures 7F and S7B), whereas displacement of BRD4 in typical enhancer regions was less profound, as has previously been reported.[23] Similar to the observed effects upon enhancer excision, we observed a profound S phase reduction along with a G0/G1 cell cycle arrest, higher rates of maturation and apoptotic events upon BRD4-inhibition (Figures 7G, S7F, S7G, and S7H). Reintroduction of EVI1 by nucleofection of MUTZ-3 cells prior to JQ1 treatment, allowing for transient *EVI1* expression, partly rescued MUTZ-3 from JQ1 cytotoxicity, arguing for relative selectivity of JQ1 for the *EVI1* super-enhancer as opposed to globally inhibiting other putative oncogenic drivers (Figure 7H).

DISCUSSION

In summary, inv(3)/t(3;3) chromosomal rearrangements cause dysregulation of two specific AML predisposition genes by aberrant activity of a single enhancer element in its ectopic chromatin environment: (I) Overexpression of *EVI1* is caused by inappropriate transcriptional control of the ectopic *GATA2* regulatory element, while (II) *GATA2* transcriptional impairment results from the removal of that same enhancer from its genomic origin. These dual events mediated by a single enhancer rearrangement, without formation of an oncogenic fusion product, highlight the vulnerability of genome organization into long-range regulatory interaction domains in case of a chromosomal break. The enhancer we identified appears to originally control transcription of the 110 kb distant *GATA2* gene at 3q21, and not the nearby gene *RPN1*. Our finding is in accordance with reports demonstrating a highly homologous -77 kb enhancer element to constitute a component of the murine *Gata2* master regulatory complex[25], and that this element is indeed leukemogenic via *EVI1* activation in transgenic mice harboring the human 3q21q26-rearranged allele.[26] In case of an inv(3)/t(3;3), the rearranged enhancer engaged in chromatin loops with the *EVI1* promoter, in certain samples over a distance of more than 200 kb. Our data emphasize that the function of an enhancer is not only determined by its location, but in particular by its ability to physically bind to an appropriate promoter, which can even occur in a different chromosome topology. Our findings show that not *RPN1,* as reported in the nomenclature of the WHO2008 classification [inv(3)/t(3;3)/*RPN1-EVI1*], but rather the *GATA2* locus is the source of the ectopic enhancer activating *EVI1* in this type of leukemia.

Besides aberrant *EVI1* activation, rewiring of parts of the *GATA2* and *EVI1* domains led to a reduction of *GATA2* expression levels. *EVI1* activation in this subtype of AML argues for a primitive HSC defect.[8,9,27-31] Since *GATA2* is a critical hematopoietic stemness factor, primitive hematopoietic precursors will be particularly susceptible to disturbances of *GATA2* homeostasis. Thus, *GATA2* deficiency may provide the right spatiotemporal context for *EVI1* oncogene activation, i.e. in the right cell at the right stage of differentiation for subsequent malignant transformation. Functional haploinsufficiency arising from inactivating mutations in GATA2 DNA-binding domains or in *GATA2* regulatory sequences represents a well-established underlying cause of MDS/AML and

Emberger/MonoMAC syndromes[32-36], of which the latter are characterized by monocytopenia, immune deficiency, and predisposition to myeloid leukemia with frequent monosomy 7. AML with inv(3)/t(3;3) most commonly associates with monosomy 7 and trilineage dysplasia, and, as demonstrated here, it is accompanied by impaired *GATA2* expression as well. It will be of particular interest to investigate whether in Emberger and MonoMAC patients 3q26 defects and consequently aberrant *EVI1* expression are also drivers of disease progression toward AML/MDS. Of note, the enhancer-containing 3q21 locus is rarely, but consistently involved in other chromosomal rearrangements with *PRDM* homologues of the *EVI1* gene (e.g. *BLIMP1/PRDM1* or *MEL1/PRDM16*) and their aberrant activation.[37] Both disease categories [inv(3)/t(3;3) and other t(3q21) AMLs] resemble each other by their high white blood cell and exceptionally high platelet counts at diagnosis. Further studies using in vivo models are warranted to investigate how the combined effects of *GATA2* haploinsufficiency and overexpression of *EVI1* or its homologues cooperate in malignant transformation of primitive hematopoietic progenitors.

The ectopic *EVI1* enhancer was embedded in a genomic region exhibiting large deposition of active chromatin marks and presence of read-through transcripts. This class of DNA elements has recently been recognized as so-called super-enhancers[23], which represent large open chromatin regions of >10 kb in size with key regulatory function for cellular identity and oncogene regulation in cancer. The observation that the *GATA2* enhancer region upon translocation had acquired characteristics of a super-enhancer, dominantly ranking in the MOLM-1 genome, provided the rationale for treatment with bromodomain/BET-inhibitors.[22] The presence of a 3q21 super-enhancer might also explain why JQ1 is effective in inv(3)/t(3;3) cell lines as opposed to various non 3q-rearranged AML cell lines with *EVI1* overexpression.[24] The effects seen after JQ1 treatment recapitulated the observations obtained by genome-editing experiments involving the translocated p300-interaction domain. Remodeling of the cancer genome by using in vivo nuclease as applied in this study helped to experimentally validate *EVI1* as an oncogenic driver lesion and warranted further pharmacologic experiments interfering with enhancer activity. These experiments emphasized that targeting *EVI1* transcriptional regulation using drugs directed against enhancer complexes could have therapeutic potential for this highly refractory subgroup of AML and diseases driven by similar mechanisms.

EXPERIMENTAL PROCEDURES

Subjects

Patient recruitment and sample processing were performed according to protocols from the German-Austrian Acute Myeloid Leukemia Study Group (AMLSG trials 06-04, 07-04, HD93A, HD98A/B) and the Dutch-Belgian Hematology/Oncology Cooperative Group (HOVON trials 04/A, 29, 42, 43, 81, 92). All studies were approved through institutional human ethics review board, and all patients provided written informed consent in accordance with the Declaration of Helsinki.

Generation of TALEN constructs

Construction of TALE DNA-binding domains directed to selected genomic loci was performed as described previously.[18] Genomic target coordinates were selected and filtered for off-target sites using the TAL Effector Nucleotide Targeter 2.0 tool (https://tale-nt.cac.cornell.edu/node/add/talen). Spacer length was defined within a range of 16-20 bp, and repeat array length was set to 20 bp. The NN repeat variable domain targeting base G was chosen in the assembly. In brief, hexamer modules were assembled from a PCR-amplified monomer library using a hierarchical digestion-ligation reaction and subsequently cloned into a full-length TALEN construct. Plasmids were verified by Sanger sequencing and tested for functionality upon transfection in HEK293T cells. To induce a genomic deletion, two TALEN pairs were transfected owing to dimerization requirement of the FokI nuclease for double-strand break formation. Repair of chromatin cleavage at the left/upstream and right/downstream boundaries of the target locus relies on non-homologous end joining (NHEJ) in the absence of a repair donor and results in the deletion of a TALEN-targeted DNA segment.

Generation of CRISPR constructs

In this study, the RNA-guided endonuclease genome-editing system was employed in experiments involving the cell line K562 owing to its cell line-specific superior targeting efficiency compared with TALENs genome-editing approaches.[38] Publicly available plasmids expressing the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)/Cas9 system were used for cloning of targeting constructs following recently published protocols.[38,39] In brief, custom target-specific oligonucleotides were cloned into a chimeric guide RNA array of an hSpCas9-expressing targeting vector. Oligonucleotides for site-specific chromatin cleavage of genomic target regions were designed following described guidelines and selected for uniqueness using a bioinformatic filtering tool (http://www.genome-engineering.org/crispr/). To induce segmental deletions of candidate regulatory DNA regions, two CRISPR plasmids were transfected into cells. Each construct was directed to flanking target site positions of the intervening DNA segment for induction of NHEJ-mediated repair upon DSB formation. Cells were screened for deletion events 48 hours later by mutation-specific PCR analogous to TALENs experiments.

Clone screening and sequencing

Upon expansion of TALEN- or CRISPR-targeted clones, genomic DNA was isolated with the QuickExtract DNA Extraction Solution (Epicentre) and screened for deletion events by mutation-specific PCR using primers spanning the breakpoint junction. A shift in amplicon size visualized by appearance of a lower running band on gel electrophoresis indicated successful targeting, and candidate clones were subsequently checked for monoclonality. The native amplicon and novel fusion fragment of candidate clones were separately purified, and sequences of informative,

heterozygous SNVs in the target region was determined by Sanger sequencing. Monoallelic targeting was confirmed by loss of heterozygosity at the SNV-specific nucleotide site. Monoclonal biallelic deletion mutants were detected by loss of the native amplicon and presence of a single, novel fusion fragment represented by the lower running band.

Acknowledgments

We thank J. Qi for providing JQ1; I.P. Touw, P. van Strien, K. Eiwen, J. van Galen, A.C. Bijkerk, P. Hogenbirk-Hupkes, J. Koenders, F. Cornelissen, H. van de Werken, and other members of the Department of Hematology at the Erasmus MC for their support; E. Simons for editorial assistance. We thank L. Cong and F. Zhang from the Broad Institute for their support with the CRISPR/Cas9 toolbox. This work was financially supported by grants from the Deutsche Forschungsgemeinschaft (GR 3955/1-1, S. Gröschel), the Lady Tata Memorial Trust, Association for International Cancer Research (E. Bindels), the Center for Translational Molecular Medicine (CTMM; GR03O-102, M. Sanders) and an EHA Research Fellowship (S. Gröschel).

REFERENCES

1. Mitelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet*. 2004;36(4):331-334.
2. Frohling S, Dohner H. Chromosomal abnormalities in cancer. *N Engl J Med*. 2008;359(7):722-734.
3. Mitelman F, Johansson B, Mertens F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2013). Mitelman F, Johansson B and Mertens F (Eds.), http://cgap.nci.nih.gov/Chromosomes/Mitelman. 2013.
4. Polack A, Feederle R, Klobeck G, Hortnagel K. Regulatory elements in the immunoglobulin kappa locus induce c-myc activation and the promoter shift in Burkitt's lymphoma cells. *EMBO J*. 1993;12(10):3913-3920.
5. Bakhshi A, Jensen JP, Goldman P, et al. Cloning the chromosomal breakpoint of t(14;18) human lymphomas: clustering around JH on chromosome 14 and near a transcriptional unit on 18. *Cell*. 1985;41(3):899-906.
6. Tsujimoto Y, Gorham J, Cossman J, Jaffe E, Croce CM. The t(14;18) chromosome translocations involved in B-cell neoplasms result from mistakes in VDJ joining. *Science*. 1985;229(4720):1390-1393.
7. Suzukawa K, Parganas E, Gajjar A, 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*. 1994;84(8):2681-2688.
8. Goyama S, Yamamoto G, Shimabe M, et al. Evi-1 is a critical regulator for hematopoietic stem cells and transformed leukemic cells. *Cell Stem Cell*. 2008;3(2):207-220.
9. Kataoka K, Sato T, Yoshimi A, et al. Evi1 is essential for hematopoietic stem cell self-renewal, and its expression marks hematopoietic cells with long-term multilineage repopulating activity. *J Exp Med*. 2011;208(12):2403-2416.
10. Pinheiro I, Margueron R, Shukeir N, et al. Prdm3 and Prdm16 are H3K9me1 methyltransferases required for mammalian heterochromatin integrity. *Cell*. 2012;150(5):948-960.
11. Morishita K, Parker DS, Mucenski ML, Jenkins NA, Copeland NG, Ihle JN. Retroviral activation of a novel gene encoding a zinc finger protein in IL-3-dependent myeloid leukemia cell lines. *Cell*. 1988;54(6):831-840.
12. Stein S, Ott MG, Schultze-Strasser S, et al. Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nat Med*. 2010;16(2):198-204.
13. Deng W, Lee J, Wang H, et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*. 2012;149(6):1233-1244.
14. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*. 2012;489(7414):109-113.
15. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012;489(7414):75-82.
16. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*. 2002;10(6):1453-1465.
17. van de Werken HJ, Landan G, Holwerda SJ, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods*. 2012;9(10):969-972.
18. Sanjana NE, Cong L, Zhou Y, Cunniff MM, Feng G, Zhang F. A transcription activator-like effector toolbox for genome engineering. *Nat Protoc*. 2012;7(1):171-192.
19. van Dongen JJ, Lhermitte L, Bottcher S, et al. EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia*. 2012;26(9):1908-1975.
20. Wieser R, Schreiner U, Rieder H, et al. Interphase fluorescence in situ hybridization assay for the detection of rearrangements of the EVI-1 locus in chromosome band 3q26 in myeloid malignancies. *Haematologica*. 2003;88(1):25-30.

21. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485(7398):376-380.

22. Loven J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*. 2013;153(2):320-334.

23. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*. 2013;153(2):307-319.

24. Zuber J, Shi J, Wang E, et al. RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature*. 2011;478(7370):524-528.

25. Grass JA, Jing H, Kim SI, et al. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol Cell Biol*. 2006;26(19):7056-7067.

26. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415-427.

27. de Pater E, Kaimakis P, Vink CS, et al. Gata2 is required for HSC generation and survival. *J Exp Med*. 2013.

28. Ling KW, Ottersbach K, van Hamburg JP, et al. GATA-2 plays two functionally distinct roles during the ontogeny of hematopoietic stem cells. *J Exp Med*. 2004;200(7):871-882.

29. Orlic D, Anderson S, Biesecker LG, Sorrentino BP, Bodine DM. Pluripotent hematopoietic stem cells contain high levels of mRNA for c-kit, GATA-2, p45 NF-E2, and c-myb and low levels or no mRNA for c-fms and the receptors for granulocyte colony-stimulating factor and interleukins 5 and 7. *Proc Natl Acad Sci U S A*. 1995;92(10):4601-4605.

30. Spinner MA, Sanchez LA, Hsu AP, et al. GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics and immunity. *Blood*. 2013.

31. Tsai FY, Orkin SH. Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood*. 1997;89(10):3636-3643.

32. Hahn CN, Chong CE, Carmichael CL, et al. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet*. 2011;43(10):1012-1017.

33. Hsu AP, Johnson KD, Falcone EL, et al. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood*. 2013;121(19):3830-3837, S3831-3837.

34. Hsu AP, Sampaio EP, Khan J, et al. Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome. *Blood*. 2011;118(10):2653-2655.

35. Ostergaard P, Simpson MA, Connell FC, et al. Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nat Genet*. 2011;43(10):929-931.

36. Rodrigues NP, Janzen V, Forkert R, et al. Haploinsufficiency of GATA-2 perturbs adult hematopoietic stem-cell homeostasis. *Blood*. 2005;106(2):477-484.

37. Lugthart S, Groschel S, Beverloo HB, 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*. 2010;28(24):3890-3898.

38. Mali P, Yang L, Esvelt KM, et al. RNA-guided human genome engineering via Cas9. *Science*. 2013;339(6121):823-826.

39. Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*. 2013;339(6121):819-823.

CHAPTER 8

Mutational spectrum of myeloid malignancies with inv(3)/t(3;3) reveals a predominant involvement of RAS/RTK signaling pathways

Stefan Gröschel[1,2,3,*], Mathijs A. Sanders[1,*], Remco Hoogenboezem[1], Annelieke Zeilemaker[1], Marije Havermans[1], Claudia Erpelinck[1], Eric M.J. Bindels[1], H. Berna Beverloo[4,5], Hartmut Döhner[3], Bob Löwenberg[1], Konstanze Döhner[3], Ruud Delwel[1,*], and Peter J. M. Valk[1,*]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands
[2] National Center for Tumor Diseases and German Cancer Research Center, Department of Translational Oncology, Heidelberg, Germany
[3] University Hospital Ulm, Department of Internal Medicine III, Ulm, Germany
[4] Erasmus University Medical Center, Department of Clinical Genetics, Rotterdam, The Netherlands
[5] Dutch Working Group on Hemato-Oncologic Genome Diagnostics

* These authors contributed equally to this work

Blood. 2015 Jan 1;125(1):133-9

ABSTRACT

Myeloid malignancies bearing chromosomal inv(3)/t(3;3) abnormalities are among the most therapy-resistant leukemias. Deregulated expression of *EVI1* is the molecular hallmark of this disease; however, the genome-wide spectrum of cooperating mutations in this disease subset has not been systematically elucidated. Here, we show that 98% of inv(3)/t(3;3) myeloid malignancies harbor mutations in genes activating RAS/receptor tyrosine kinase (RTK) signaling pathways. In addition, hemizygous mutations in *GATA2*, as well as heterozygous alterations in *RUNX1*, *SF3B1*, and genes encoding epigenetic modifiers frequently co-occur with the inv(3)/t(3;3) aberration. Notably, neither mutational patterns nor gene expression profiles differ across inv(3)/t(3;3) AML, CML, and MDS cases, suggesting recognition of inv(3)/t(3;3) myeloid malignancies as a single disease entity irrespective of blast count. The high incidence of activating RAS/RTK signaling mutations may provide a target for a rational treatment strategy in this high-risk patient group.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter8/

INTRODUCTION

Acute myeloid leukemia (AML) with inv(3)(q21q26.2) or t(3;3)(q21;q26.2) [inv(3)/t(3;3)] is a distinct disease entity in the current World Health Organization classification.[1] High therapy-resistance is the common feature of myeloid malignancies, particularly AML with 3q21/3q26 aberrations, manifesting in low rates of complete remission and subsequent failure of current treatment strategies.[2-4] Appearance of the characteristic 3q aberrations also indicates disease progression and portends adverse outcome in myelodysplastic syndrome (MDS) and chronic myeloid leukemia (CML).[5-7] Therapy resistance in this subtype of malignancies is linked to the inappropriate activation of the proto-oncogene *Ecotropic Viral Integration-1* (*EVI1*) as a consequence of the chromosome 3 rearrangements. EVI1 is a hematopoietic stemness factor and transcription factor with chromatin remodeling activity.[8-10] *EVI1* is also overexpressed in approximately 11% of all AML cases in the absence of 3q aberrations and represents an independent adverse prognostic factor in these patients.[11] We and others have shown that, as a consequence of inv(3)/t(3;3) rearrangements, *EVI1* becomes activated via structural repositioning of a distal *GATA2* enhancer from 3q21 to the *EVI1* locus at 3q26.[12,13] Relocation of the enhancer additionally confers reduced and monoallelic *GATA2* expression in this AML subtype. Notably, GATA2 deficiency has been shown to impair hematopoietic stem cell frequency and fitness[14-16], and *Evi1* activation in murine inv(3)/t(3;3) models is followed by leukemia onset after a long latency of 6 months.[13] Hence, we hypothesize that additional cooperating genetic events, other than *EVI1* and *GATA2* deregulation, are required for full leukemic transformation, resulting in a myeloid disease with dismal outcome. Full understanding of the complete spectrum of molecular defects associated with this highly refractory AML subtype may provide additional rationale for treatment and to overcome therapeutic nihilism in this incurable disease category. Therefore, within this study, we sought to extend the molecular characterization of myeloid disorders with inv(3)/t(3;3) aberrations by next-generation sequencing (NGS).

METHODS

Patient samples

From the combined study groups of the Dutch-Belgian Cooperative Trial Group for Hematology-Oncology (HOVON) and the German-Austrian AML Study Group (AMLSG) we selected 32 AML (including 2 cell lines MUTZ-3 and UCSD-AML1), 4 CML-BC (including 2 cell lines HNT-34 and MOLM-1), and 5 MDS cases for NGS analysis. Included patients harbored an inv(3)/t(3;3) aberration on chromosome banding analysis (Supplemental Table 1) and was subsequently confirmed by NGS analysis. Cultured $CD3^+$ T-cells from diagnostic bone marrow served as whole exome sequencing (WES) germline control. Written informed consent was obtained from all individuals. All samples were sequenced on the Illumina HiSeq 2500 system and processed as described previously.[12]

3q-capture sequencing

From the collected patient material the genomic DNA was sheared with the Covaris S2 device (Covaris) with default settings. Subsequently, the sample libraries were prepared using the TruSeq DNA Sample Preparation Guide (Illumina). The target chromosomal regions 3q21.1-3q26.2 (~40Mb) were captured by employing a custom in-solution oligonucleotide baits (Nimblegen SeqCap EZ Choice XL). The final sample libraries were subjected to paired-end sequencing (2x100bp) and were aligned against the human genome 19 (hg19) using the Burrows Wheeler Aligner (BWA) with default settings.[17] Exact breakpoint positions were determined with Breakdancer v1.1.[18] Exact breakpoint sequences were resolved by extracting proximal reads supporting or spanning the breakpoint using the identified breakpoint positions and an algorithm able to extract the relevant reads from BAM files by employing the Samtools API.[19] Relevant reads were identified by their discordant distance to the paired mate read due to the inv(3)/t(3;3) aberration (supporting reads) or being a member of a cluster of truncated reads with the same clipping position (spanning reads). The extracted reads were subsequently used as input for the *de novo* assembler Velvet v1.0.17[20] with default settings and the assembled region was validated with UCSC Blat.[21] If resolved, the breakpoint sequences of 3q21 and 3q26 were used for the estimation of the variant allele frequency (VAF) to infer the cellular prevalence of the inv(3)/t(3;3) aberration. All 3q-capture sequencing (3q-Seq) reads were aligned against the resolved breakpoint sequences of 3q21 and 3q26 and their respective native wild type sequences. The VAF was estimated by comparing the total number of reads aligning on the breakpoint sequence to the total number of reads aligning to the respective native wild-type sequence.

RNA-Seq and whole exome sequencing

From the collected patient material total RNA was extracted with phenol-chloroform and subsequently transcribed by using Superscript II RT (Invitrogen). Shearing of the cDNA was performed with the Covaris S2 device (Covaris) with the default settings and further constructed according to the TruSeq RNA Sample Preparation v2 Guide (Illumina). The sample libraries were subjected to paired-end sequencing (2x75bp) and aligned against hg19 using TopHat v2.[22] Genomic DNA from patients and *in vitro* cultured control $CD3^+$ T-cells were processed similar to 3q-Seq protocols and captured by exome bead capture (SeqCap EZ Human Exome Library v3.0). The sample libraries were paired-end sequenced (2x100bp) and subsequently aligned against hg19 using BWA with default settings.[17]

Overall, we performed whole transcriptome sequencing (RNA-Seq) on 41 and WES on 10 out of these 41 inv(3)/t(3;3) myeloid malignancies. Read and alignment statistics for RNA-Seq and WES data are found in the Supplement (Supplemental Figures 1A-C, Supplemental Table 4). On average we observed a medium to high coverage for the targeted exome in WES data (~62x), sufficient to detect mutations with a VAF of 10% or more. Reads generated for RNA-Seq analyses

predominately fell within transcribed regions (~52%), i.e. ribosomal genes, coding sequence, and UTRs, according to the RefSeq Transcriptome database, and on average 91% of the reads could be aligned to hg19. Gene expression profiles (GEP) for 24 inv(3)/t(3;3) patients were constructed for differential expression, clustering, and principle component analyses with the DESeq2 package.[23] Copy number variation (CNV) profiles from the WES data were calculated by CNVsvd (M.A.S., R.H, and P.J.M.V., manuscript in preparation; Supplemental Figure 2). In brief, per patient the total number of fragments was determined for each exon or determined from consecutive 500 nucleotide-wide windows for large exons. The estimation of CNVs is hampered by systematic variance introduced by sequence technology bias or repetitive and homologous sequences, which can be observed in all sequenced cases. By utilizing a control reference data set under the assumption that these cases have a normal karyotype (i.e., the *in vitro c*ultured $CD3^+$ T-cells) allows for the determination of the local variance composition. These estimated local variance components can be used to attenuate the systematic variance in all sequenced cases. Finally, the normalized count statistics were used for the estimation of the CNV WES profile.

Variant detection

RNA-Seq data were preprocessed for variant detection by splitting the exon boundary spanning reads using the Genome Analysis Toolkit (GATK).[24] Subsequently, the variants were determined with the Samtools API and MuTect for RNA-Seq and WES data.[19,25] The detected variants were annotated with AnnoVar[26] and further characterized by multiple read statistics determined by an in-house developed algorithm. In brief, the algorithm determines for each variant the VAF, local read statistics based on the alignment and base qualities, mutation likelihood given the local sequence context, recurrence given the catalogue of somatic mutations in cancer (COSMIC), recurrence determined from population-based sequencing efforts (1000 genomes project), and, when available, the likelihood of the mutation given the same set of read statistics in a control sample. The validity of our approach combining WES data with RNA-Seq data to infer variants is substantiated by the observation that nonsense-mediate decay (NMD) was negligible for mutant allele detection, as demonstrated by similar VAFs of mutant disease alleles observed within cases characterized by both WES and RNA-Seq (Supplemental Figure 3). Frameshift and premature stop codon-introducing mutations were selected and dichotomized on their location in the gene body. Mutations located in the terminal exon or approximately 50 bp from the exon boundary of the penultimate exon should, theoretically, be unaffected by NMD while stop codon-introducing mutations situated in other locations of the gene body should be affected. Finally, variants were examined when recurrently detected in more than two patients or previously linked to leukemogenesis or cancer pathogenesis.[27,28] All listed variants were validated by Sanger sequencing, except for *FLT3-ITD* which was determined by RT-PCR.

Allelic imbalance of GATA2

In total, 30 inv(3)/t(3;3) cases accommodate informative heterozygous single nucleotide variants (SNV) in the *GATA2* locus according to the 3q-Seq data. We have previously shown that the inv(3)/t(3;3) causes monoallelic expression of *GATA2* from the non-rearranged allele.[12] Subsequently, we determined the allelic contribution of the genotypes of the heterozygous SNV in the matched RNA-Seq case. The average of the allelic contribution was taken when multiple heterozygous SNVs were accommodated in the *GATA2* locus. The polar histogram was constructed with the R package "phenotypicForest".[29]

Clonality analysis

The VAFs of the acquired mutations were estimated from the 10 paired inv(3)/t(3;3) myeloid malignancies characterized with WES. The VAF of the inv(3)/t(3;3) aberration was estimated from the 3q-Seq data, unless the breakpoints could not be resolved or no 3q-Seq data was available. In these cases the cytogenetically determined inv(3)/t(3;3) positive metaphases were used. The VAFs were corrected by the local CNV, determined by CNVsvd, and possible loss-of-heterozygosity, ascertained by determining the loss of proximal heterozygous SNVs with respect to the control WES data. The clonal architecture was illustrated in violin plots. In brief, the density of mutations with a similar VAF was determined by a kernel density approach and is represented by the width of the graph. These plots were generated by the R package "easyGgplot2".[30]

RESULTS

Mutant disease allele categorization

We first assigned mutations to mutational categories in order to discern patterns of mutations within inv(3)/t(3;3) myeloid disease (Figure 1A).[28] All identified mutations were confirmed to be somatic in samples with available paired T-cell control (10 out of 41 cases). In addition to the "hardwired" deregulated expression of *EVI1* and *GATA2*, all 41 samples contained at least one additional mutation in one of the categories relevant for leukemia pathogenesis (average 2.3 category mutations per sample [Figure 1A, Supplemental Tables 2 and 3]). Notably, all AML and CML-BC, as well as 4 of 5 MDS samples contained mutations in genes activating RAS/RTK signaling, amounting to an incidence of 98% of all malignancies with an inv(3)/t(3;3). Furthermore, mutations were frequently found in myeloid transcription factor genes (32%), splice factor-encoding genes (29%), epigenetic modifier genes (29%), tumor-suppressor genes (10%), DNA-methylation genes (10%), and cohesin-complex genes (5%) (Figure 1A).

Complementing previous reports on the high incidence of *NRAS* mutations in inv(3)/t(3;3) AML,[3,6] we found on aggregate 47% of all samples containing mutations directly affecting RAS, i.e. *NRAS* (27%), *KRAS* (11%), and *NF1* (9%) (Figure 1B). These mutations were mutually exclusive and

also largely non-overlapping with any other mutation affecting signaling pathways involving RAS, i.e. *PTPN11* (20%), *FLT3* (13%), *CBL* (7%), *KIT* (2%), and *BCR-ABL1* (12%) (Figure 1B). *GATA2* was the most commonly mutated transcription factor in inv(3)/t(3;3) myeloid malignancies (15%; 5 AML and 1 MDS patient) and occurred in all cases in one of the two GATA2 zinc finger domains. *RUNX1* mutations were present in 12% and did not coincide with *GATA2* mutations, however mutations in the splice factor encoding gene *SF3B1* (27%) were enriched in *GATA2* mutated samples. Mutations in *GATA2*, *SF3B1*, and *RUNX1* were established to be somatic in all cases with control material available. Interestingly, we detected novel truncating mutations and CNVs resulting in the loss of one copy of the transcription factor *FOXP1* (10%), which is recurrently involved in chromosomal aberrations within lymphoma[31], but its association with AML pathogenesis is unknown. The predominant monosomal karyotype within inv(3)/t(3;3) myeloid malignancies, mainly conferred by monosomy 7 (68%), is contrasted by the low incidence of *TP53* mutations (5%) (Figure 1A), which had been suggested to be involved in the etiology of complex and of monosomal karyotype AML.[32,33]

No mutational pattern alluded to the high coincidence of the loss of chromosome 7 in inv(3)/t(3;3) myeloid disease (Figure 1A**,** Supplemental Tables 2 and 3). However, previous reports have indicated that haploinsufficiency for *CUX1*, a gene strongly downregulated in our cohort of inv(3)/t(3;3)/-7 patients (Supplemental Table 5), activates phosphoinositide 3-kinase (PI3K) signaling by transcriptional downregulation of the PI3K inhibitor PI3KIP1,[34] and could therefore be an important cooperating lesion in inv(3)/t(3;3)/monosomy 7 myeloid syndromes.[35]

To date, no independent prognostic factor within the inv(3)/t(3;3) AML subset has been identified due to its low incidence and the extremely short median survival of inv(3)/t(3;3) AML patients (10 months).[3] Baseline patient characteristics and clinical outcome data were available in 21 individuals with inv(3)/t(3;3) AML. The high frequency of RAS/RTK pathway mutations allowed us to perform an exploratory analysis within this small patient cohort. There were no statistically significant differences in patient characteristics, nor overall and event-free survival in cases with RAS mutations (*NRAS*^{mut}, *KRAS*^{mut}, *NF1*^{mut}) compared to cases with other mutations activating signaling pathways (Supplemental Figure 4). The median overall survival (OS) of RAS^{mut} patients was 9.8 months versus 8.9 months of RTK^{mut} patients.

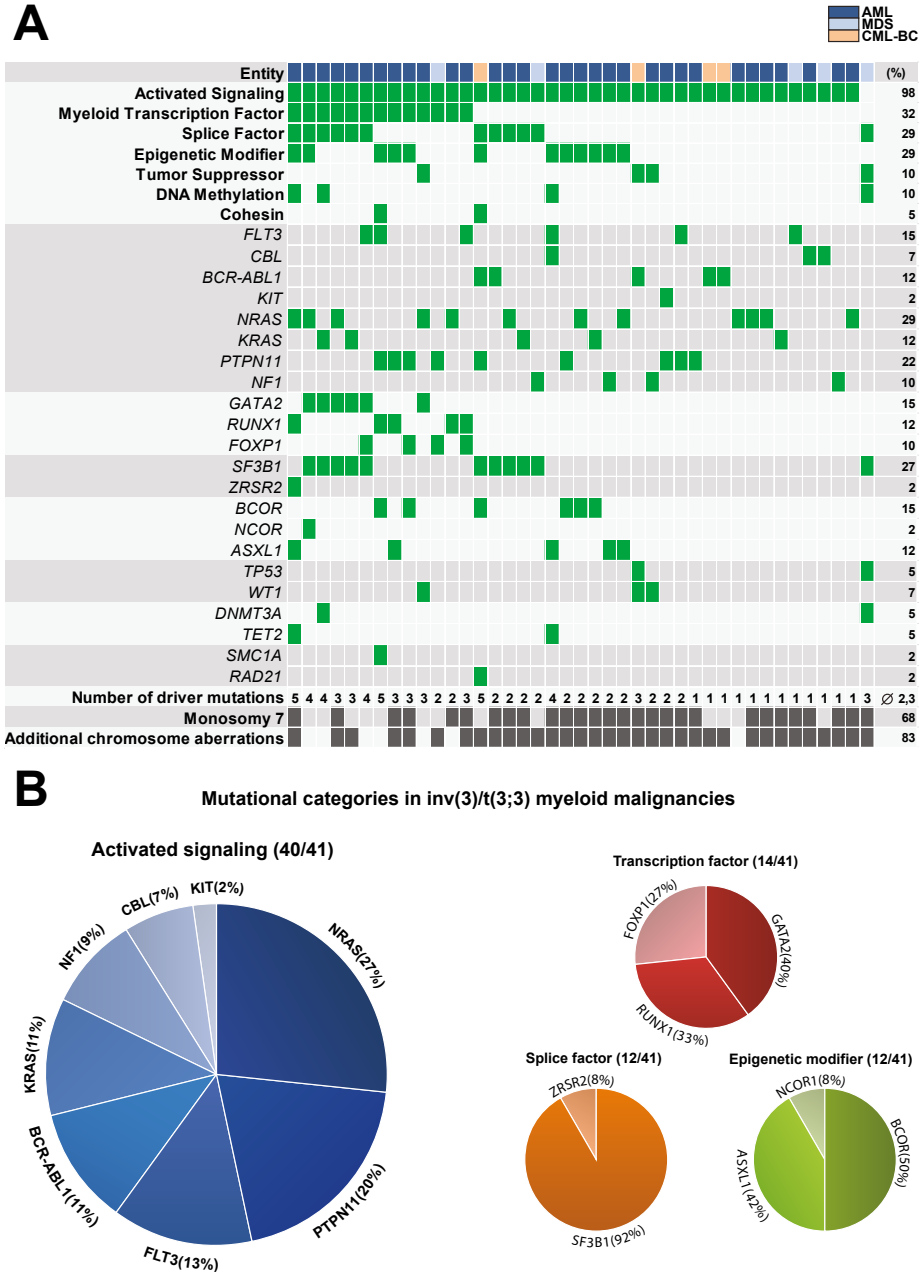


Figure 1. Mutational spectrum observed in inv(3)/t(3;3) myeloid malignancies. (A) Distribution of acquired mutations in 41 inv(3)/t(3;3) myeloid malignancies in conjunction with the aggregation of the mutations into mutational categories: mutant (green), positive for cytogenetic abnormality (dark gray), wild-type (gray). (B) Distribution of mutations within mutational categories present in more than 10% of the inv(3)/t(3;3) myeloid malignancies.

Clonality analysis

To address the question whether the highly overrepresented RAS/RTK pathway mutations and other recurrent somatic alterations in inv(3)/t(3;3) AML co-occurred in the same dominant clone, we assessed the allelic ratios of the *EVI1*-rearranged and mutant candidate disease alleles (Figure 2). WES analysis in conjunction with germline T-cell control was available from 10 AML patients. Cytogenetic evaluation of blast percentage and NGS read count estimation of the percentage of the 3q21q26.2 fusion (allele frequency) were concordant. In two cases (AML 20908 and 29656) without available 3q-Seq data cytogenetics served to estimate the percentage of the inv(3) allele. The inv(3)/t(3;3) aberrations were detectable in the majority of cases (7/10 cases) in up to 100% of the cells (i.e., resulting in an allelic ratio of the heterozygous 3q21q26.2 fusion allele of approximately 0.5), reflecting high blast percentage in these cases. The RAS and RTK mutations were mainly found in the dominant *EVI1*-rearranged clone, and a similar pattern is found for all other identified alterations (e.g., in transcription factor, splice factor, and epigenetic modifier genes), which mostly co-occur at similar frequency as the RAS/RTK mutations.

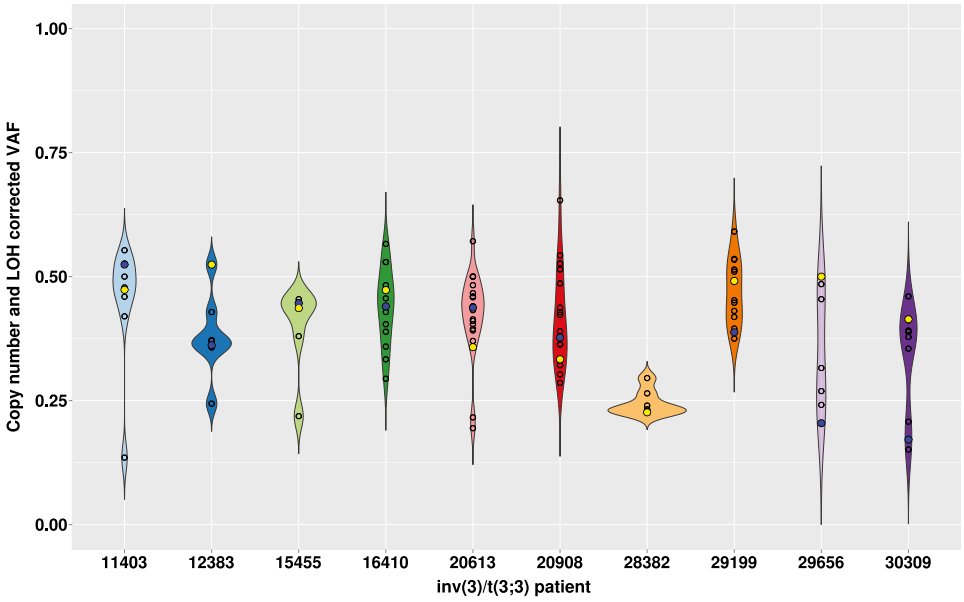


Figure 2. Clonal architecture inferred from somatic mutations observed in 10 inv(3)/t(3;3) myeloid malignancies. Distribution of estimated variant allele frequencies (VAF) determined from whole exome sequencing. The width of the graph represents the density of mutations with similar VAFs. The yellow dot denotes the VAF of the 3q-aberration [(inv(3)/t(3;3)], the blue dot denotes the VAF of the RAS/RTK-associated mutation, and open circles denote the VAF of all other mutations.

However, in AML 12383 (*PTPN11* mutation) and AMLs 29656 and 30309 (both *NF1*-mutated), the 3q-rearrangement was found in the major clone, whereas the RAS pathway mutations were

present in only about half of these cells. In 2 cases (AML 20613 and 20908) the inv(3)/t(3;3) aberrations were less frequent than other concomitant mutations. In the inv(3) MDS case 28382 without any detected activating signaling mutation, the allelic ratio of the inv(3) was about 0.25, suggesting that both dysplastic-appearing cells as well as myeloblasts (blast percentage as per cytologic evaluation <20%) carried both the inv(3) aberration and coincident gene mutations (*SF3B1*, *TP53*, *DNMT3A*; see Figure 1A). Together, these data suggest that the inv(3) or t(3;3) aberration is the primary genetic hit in this subset of malignancies, with high proportion of clones harboring concurrent activating signaling mutations. Owing to the very short survival of these patients and general failure to achieve CR, no time-course monitoring could be performed to reveal clonal evolution.

Expression of mutant *GATA2*

The inv(3)/t(3;3) chromosomal rearrangements separate an upstream *GATA2* enhancer from 3q21 and fuse it to the 3q26.2/*EVI1* locus, thereby acquiring features of a monoallelic super-enhancer on the rearranged 3q allele.[12,13,36,37] Integrative analysis of RNA-seq with 3q-capture DNA-seq data using informative, heterozygous SNPs revealed almost exclusive monoallelic expression of the mutant *GATA2* alleles in 30 inv(3)/t(3;3) cases including cell lines available for analysis (Figure 3), as shown in the polar plot by the contribution of the rearranged 3q and nonrearranged 3q allele read counts for *GATA2* on the basis of the SNVs (SNPs plus somatic mutations) for each patient.

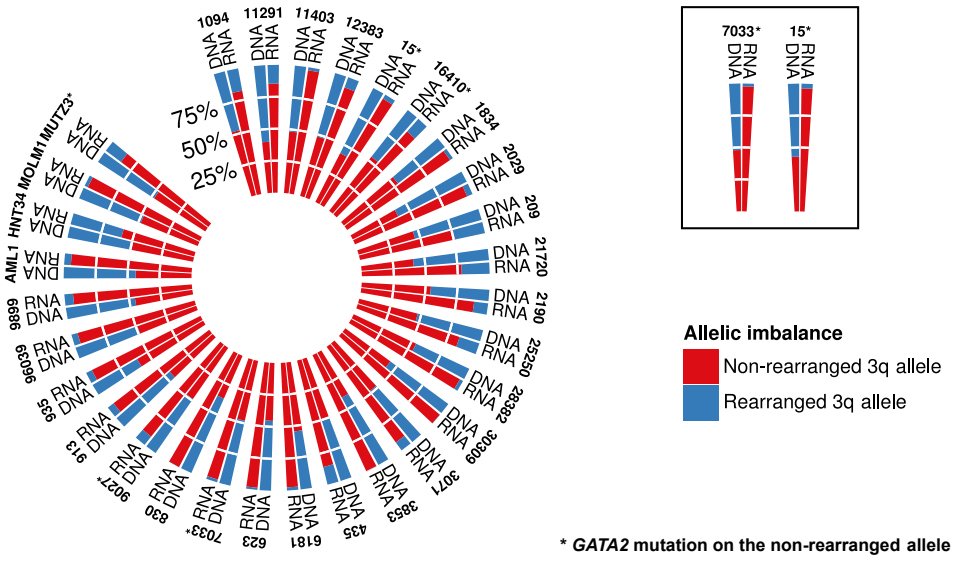


Figure 3. Polar histogram plot of allelic imbalance of *GATA2* expression observed in RNA-Seq. For each patient the average VAF is estimated from informative heterozygous SNVs from the 3q-Seq data. Average RNA-Seq *GATA2* VAF is estimated from the same SNV positions. Asterisks denote the presence of a somatic *GATA2* mutation in the indicated sample on the non-rearranged allele.

This observation indicates that the remaining active, non-rearranged *GATA2* allele acquired the mutation, whereas the non-mutated *GATA2* allele was silenced due to the chromosomal rearrangement. Thus, in our inv(3)/t(3;3) AML cohort, heterozygous *GATA2* mutations were "functionally" hemizygous due to *GATA2* silencing.

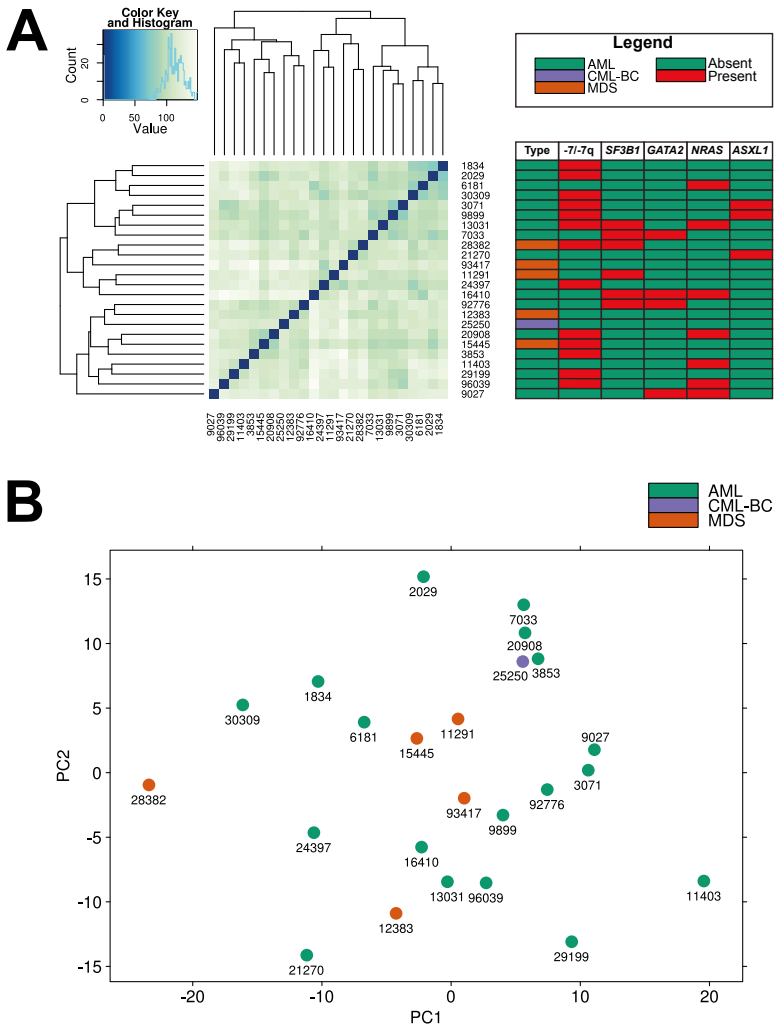


Figure 4. Gene expression profile (GEP) analysis. (A) Clustering of GEPs using the Euclidean distance metric and hierarchical clustering. The adjacent table details the acquired mutations of the clustered malignancies. (B) Principle component analysis performed on the GEPs of 24 inv(3)/t(3;3) myeloid malignancies displays no discrimination between AML and MDS.

Gene expression and mutation patterns in AML and MDS

It is a matter of debate whether MDS with the distinct inv(3)(q21q26.2) or t(3;3)(q21;q26.2) should be regarded as AML irrespective of blast percentage in the bone marrow, similar to the current WHO guidelines applied in diagnosis of core binding factor AML with inv(16)/t(16;16) or t(8;21) and of acute promyelocytic leukemia with t(15;17).[1,5,6,38] In an effort to discriminate MDS and AML with inv(3)/t(3;3) based on gene expression programs and the spectrum of coincident gene mutations we performed cluster and principle component analyses (Figures 4A and 4B). No cluster formation emerged, neither based on the MDS/AML dichotomy nor any other unsubstantiated group within our dataset. Furthermore, we performed a differential expression analysis to infer genes that could differentiate between MDS and AML. In summary, after Benjamini-Hochberg correction for multiple testing, we could only detect two differentially expressed genes (*C11orf45*: p=0.0009, *CILP*: p=0.04) without a documented role in leukemogenesis. Additionally, we observed that MDS patients with inv(3)/t(3;3) are equally therapy-resistant as their AML counterparts in a small set of cases analyzed (data not shown). In conclusion, we were unable to detect cluster formation, indicating the strong homogeneity of inv(3)/t(3;3) myeloid malignancies based on GEPs and the pattern of cooperating genetic lesions.

DISCUSSION

Collectively, we present data that suggest a common genetic background of myeloid malignancies harboring inv(3) or t(3;3) and show that RAS alterations and activating RTK mutations are more frequent in this disease subset than previously reported.[3,6,39,40] The spectrum of secondary genetic lesions is generally found in the same *EVI1*-rearranged dominant clone. No unique cluster within inv(3)/t(3;3) myeloid malignancies could be identified, neither by gene expression or mutation profiling, nor by analysis of patient characteristics or clinical outcome. Thus, our data further support the notion that inv(3)/t(3;3) myeloid disorders could be categorized as AML irrespective of blast count, similar to WHO AML categories t(8;21), inv(16)/t(16;16), or t(15;17), which is also suggested by the molecular pathobiology common to all inv(3)/t(3;3) myeloid malignancies.[12,13]

Reclassification of the currently annotated WHO AML subtype inv(3)/t(3;3); *RPN1-EVI1* as inv(3)/t(3;3); *GATA2*-EVI1-rearranged AML is supported by the observation that *GATA2* allelic imbalances and monoallelic expression of heterozygous *GATA2* mutations occur due to the distinct chromosomal rearrangements. Whether or not this and other myeloid transcription factor alterations contribute to disease biology and the highly adverse clinical phenotype of inv(3)/t(3;3) patients remains to be shown, although *GATA2* and other transcription factor disturbances have described to be preleukemic lesions.[28,41-45] Of note, myeloid malignancies with inv(3) or t(3;3) define yet another subset of AML with high enrichment of *GATA2* mutations next to *CEBPA*-mutated AML.[46,47]

We included CML cases in blast crisis with an inv(3)/t(3;3) under the assumption that CML-BC closely resembles AML biology.[48] The *BCR-ABL1* fusion is an RTK mutant that in itself activates RAS

pathways and is the first event in transformation of myeloid precursors, as opposed to MDS and AML cells that first acquired inv(3)/t(3;3).[49,50] Despite the difference of biology and etiology of CML, the mutational spectrum of inv(3)/t(3;3) CML-BC cells appears to be same, as further suggested by transcriptome analysis, which showed that gene expression profile of the single CML-BC case did not differ from that of AML and MDS cases. However, the small number of inv(3)/t(3;3) MDS and CML cases in our study preclude conclusive assessment of the role of inv(3)/t(3;3) with regards to clinical phenotype.

In summary, inv(3)/t(3;3) myeloid malignancies harbor a common set of molecular alterations, i.e. *EVI1* and *GATA2* deregulation coupled with mutations activating key signaling pathways. The dependence on constitutive RAS/RTK signaling activity of inv(3)/t(3;3) transformed AML cells might be the molecular correlate of the observed high white blood cell counts in this disease subset. Also, in view of the negative impact of *GATA2* deficiencies on proliferation and regeneration of myeloid progenitors,[15,41,51,52] these activated signaling mutations may be indispensable for survival and propagation of inv(3)/t(3;3)-transformed myeloid progenitors. The high mutational burden of inv(3)/t(3;3) cells as compared to other AML subtypes[27] (Supplemental Table 2) could also provide clues as to why inv(3)/t(3;3) malignancies invariably associate with an extremely poor prognosis. As these rare inv(3)/t(3;3) myeloid malignancies form a highly unmet medical need, novel therapeutic approaches could be derived from the observation of constitutive activation of the MAPK pathway in almost 100% of these tumors. Exploiting signaling pathways therapeutically by using FLT3- or PI3K-inhibitors[53] or hypothetically by interfering with RAS-signaling, possibly in combination with BET-inhibitors[12], may serve as valuable adjuncts to the scarce armamentarium of chemotherapeutic drugs effective in this subset of malignancies.

Acknowledgements

This work was financially supported by grants from the Deutsche Forschungsgemeinschaft (GR3955/1-1, S.G.), the Lady Tata Memorial Trust (S.G.), the Center for Translational Molecular Medicine (CTMM; GR03O-102, M.A.S.), an EHA Research Fellowship (S.G.) and the Worldwide Cancer Research (formerly AICR) (12-1309) (E.M.J.B.).

Authorship contributions

S.G., M.A.S., R.D., and P.J.M.V. designed research, performed experiments, analyzed and interpreted data, and wrote the manuscript; S.G., A.Z., M.H., R.H., E.B., C.E. generated NGS libraries and performed Sanger and Illumina sequencing; H.B.B., B.L., K.D., H.D., and P.J.M.V. collected specimens and clinical data; H.B.B., K.D., H.D., and P.J.M.V. performed cytogenetic and molecular analyses of leukemia samples.

Disclosure of conflicts of interest

The authors declare no competing conflicts of interest.

REFERENCES

1. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*. 2009;114(5):937-951.
2. Grigg AP, Gascoyne RD, Phillips GL, Horsman DE. Clinical, haematological and cytogenetic features in 24 patients with structural rearrangements of the Q arm of chromosome 3. *Br J Haematol*. 1993;83(1):158-165.
3. Lugthart S, Groschel S, Beverloo HB, 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*. 2010;28(24):3890-3898.
4. Testoni N, Borsaru G, Martinelli G, et al. 3q21 and 3q26 cytogenetic abnormalities in acute myeloblastic leukemia: biological and clinical features. *Haematologica*. 1999;84(8):690-694.
5. Cui W, Sun J, Cotta CV, Medeiros LJ, Lin P. Myelodysplastic syndrome with inv(3)(q21q26.2) or t(3;3)(q21;q26.2) has a high risk for progression to acute myeloid leukemia. *Am J Clin Pathol*. 2011;136(2):282-288.
6. Haferlach C, Bacher U, Haferlach T, et al. The inv(3)(q21q26)/t(3;3)(q21;q26) is frequently accompanied by alterations of the RUNX1, KRAS and NRAS and NF1 genes and mediates adverse prognosis both in MDS and in AML: a study in 39 cases of MDS or AML. *Leukemia*. 2011;25(5):874-877.
7. Johansson B, Fioretos T, Mitelman F. Cytogenetic and molecular genetic evolution of chronic myeloid leukemia. *Acta Haematol*. 2002;107(2):76-94.
8. Cattaneo F, Nucifora G. EVI1 recruits the histone methyltransferase SUV39H1 for transcription repression. *J Cell Biochem*. 2008;105(2):344-352.
9. Goyama S, Yamamoto G, Shimabe M, et al. Evi-1 is a critical regulator for hematopoietic stem cells and transformed leukemic cells. *Cell Stem Cell*. 2008;3(2):207-220.
10. Perkins AS, Fishel R, Jenkins NA, Copeland NG. Evi-1, a murine zinc finger proto-oncogene, encodes a sequence-specific DNA-binding protein. *Mol Cell Biol*. 1991;11(5):2665-2674.
11. Groschel S, Lugthart S, Schlenk RF, et al. High EVI1 expression predicts outcome in younger adult patients with acute myeloid leukemia and is associated with distinct cytogenetic abnormalities. *J Clin Oncol*. 2010;28(12):2101-2107.
12. Groschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369-381.
13. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415-427.
14. Hsu AP, Johnson KD, Falcone EL, et al. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood*. 2013;121(19):3830-3837, S3831-3837.
15. Lim KC, Hosoya T, Brandt W, et al. Conditional Gata2 inactivation results in HSC loss and lymphatic mispatterning. *J Clin Invest*. 2012;122(10):3705-3717.
16. Ling KW, Ottersbach K, van Hamburg JP, et al. GATA-2 plays two functionally distinct roles during the ontogeny of hematopoietic stem cells. *J Exp Med*. 2004;200(7):871-882.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
18. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6(9):677-681.
19. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
20. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 2008;18(5):821-829.
21. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656-664.

22. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.

23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*. 2014.

24. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.

25. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219.

26. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.

27. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013;502(7471):333-339.

28. TCGA. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368:2059-2074.

29. Ladroue C. http://chrisladroue.com/phorest/.

30. Kassambara A. http://www.sthda.com/english/articles/3-easyggplot2/.

31. Streubel B, Vinatzer U, Lamprecht A, Raderer M, Chott A. T(3;14)(p14.1;q32) involving IGH and FOXP1 is a novel recurrent chromosomal aberration in MALT lymphoma. *Leukemia*. 2005;19(4):652-658.

32. Breems DA, Van Putten WL, De Greef GE, et al. Monosomal karyotype in acute myeloid leukemia: a better indicator of poor prognosis than a complex karyotype. *J Clin Oncol*. 2008;26(29):4791-4797.

33. Rucker FG, Schlenk RF, Bullinger L, et al. TP53 alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome. *Blood*. 2012;119(9):2114-2121.

34. Wong CC, Martincorena I, Rust AG, et al. Inactivating CUX1 mutations promote tumorigenesis. *Nat Genet*. 2014;46(1):33-38.

35. McNerney ME, Brown CD, Wang X, et al. CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. *Blood*. 2013;121(6):975-983.

36. Grass JA, Jing H, Kim SI, et al. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol Cell Biol*. 2006;26(19):7056-7067.

37. Koche RP, Armstrong SA. Genomic dark matter sheds light on EVI1-driven leukemia. *Cancer Cell*. 2014;25(4):407-408.

38. Dohner H, Estey EH, Amadori S, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*. 2010;115(3):453-474.

39. Renneville A, Roumier C, Biggio V, et al. Cooperating gene mutations in acute myeloid leukemia: a review of the literature. *Leukemia*. 2008;22(5):915-931.

40. Tsurumi S, Nakamura Y, Maki K, et al. N-ras and p53 gene mutations in Japanese patients with myeloproliferative disorders. *Am J Hematol*. 2002;71(2):131-133.

41. de Pater E, Kaimakis P, Vink CS, et al. Gata2 is required for HSC generation and survival. *J Exp Med*. 2013;210(13):2843-2850.

42. Hahn CN, Chong CE, Carmichael CL, et al. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet*. 2011;43(10):1012-1017.

43. Kurokawa M, Nishimoto N, Arai S, et al. Loss of AML1/Runx1 accelerates the development of MLL-ENL leukemia through down-regulation of p19(ARF). *Blood*. 2011;118(9):2541-2550.

44. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med*. 2009;361(11):1058-1066.

45. Watanabe-Okochi N, Kitaura J, Ono R, et al. AML1 mutations induced MDS and MDS/AML in a mouse BMT model. *Blood*. 2008;111(8):4297-4308.

46. Green CL, Tawana K, Hills RK, et al. GATA2 mutations in sporadic and familial acute myeloid leukaemia patients with CEBPA mutations. *Br J Haematol*. 2013;161(5):701-705.

47. Greif PA, Dufour A, Konstandin NP, et al. GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. *Blood*. 2012;120(2):395-403.

48. Radich JP. The Biology of CML blast crisis. *Hematology Am Soc Hematol Educ Program*. 2007:384-391.

49. Calabretta B, Perrotti D. The biology of CML blast crisis. *Blood*. 2004;103(11):4010-4022.

50. Sawyers CL, McLaughlin J, Witte ON. Genetic requirement for Ras in the transformation of fibroblasts and hematopoietic cells by the Bcr-Abl oncogene. *J Exp Med*. 1995;181(1):307-313.

51. Hsu AP, Sampaio EP, Khan J, et al. Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome. *Blood*. 2011;118(10):2653-2655.

52. Rodrigues NP, Janzen V, Forkert R, et al. Haploinsufficiency of GATA-2 perturbs adult hematopoietic stem-cell homeostasis. *Blood*. 2005;106(2):477-484.

53. Park S, Chapuis N, Tamburini J, et al. Role of the PI3K/AKT and mTOR signaling pathways in acute myeloid leukemia. *Haematologica*. 2010;95(5):819-828.

CHAPTER 9

RNA sequencing reveals a unique fusion of the lysine (K)-specific methyltransferase 2A and smooth muscle myosin heavy chain 11 in myelodysplastic syndrome and acute myeloid leukemia

Mathijs A. Sanders[1], François G. Kavelaars[1], Annelieke Zeilemaker[1], Adil S.A. Al Hinai[1], Saman Abbas[1], H. Berna Beverloo[2], Kirsten van Lom[1] and Peter J.M. Valk[1]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands

[2] Erasmus University Medical Center, Department of Clinical Genetics, Rotterdam, The Netherlands

Haematologica. 2015 Jan;100(1):e1-e3l

ABSTRACT

Myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML) are heterogeneous malignancies characterized by a variety of acquired genetic abnormalities and variable response to treatment.[1,2] In the last decade a number of novel molecular genetic abnormalities have been revealed in MDS and AML by applying novel genome-wide technologies, such as massively parallel sequencing.[2,3] The different recurrent genetic aberrations shed light on possible mechanisms involved in leukemogenesis and refine risk-stratification of both diseases.[1] Although recurrence of aberrations in MDS and AML is the major guide to reveal general mechanisms regarding leukemogenesis, unique abnormalities can also be highly informative. Here we describe a unique fusion of the lysine (K)-specific methyltransferase 2A (*KMT2A*) gene (mixed-lineage leukemia gene (*MLL*)), located on chromosome 11q23, and the gene encoding smooth muscle myosin heavy chain 11 (*MYH11*), located on chromosome 16p13, in a patient with MDS and subsequently AML, both harboring the cryptic translocation t(11;16). *KMT2A* and *MYH11* are involved in recurrent translocations in AML, but fusions of these two genes have never been demonstrated.

INTRODUCTION

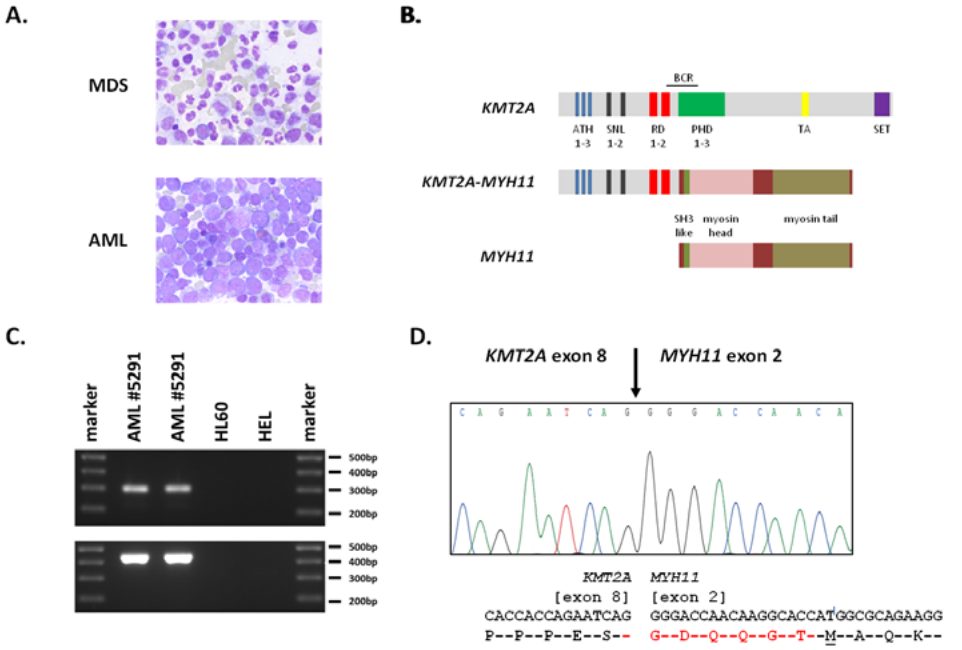
KMT2A is a transcriptional activator, which regulates gene expression, including *HOX* genes, by methylation of histone H3 lysine 4 (H3K4).[4] The *KMT2A* gene on 11q23 is involved in translocations in approximately 5% of adult AML cases[4] and more than 70 translocation partners of *KMT2A* have been described.[5] The majority of *KMT2A* fusions incorporate the N-terminal portion of KMT2A, containing three short AT-hook motifs, two speckled nuclear localization sites and a transcriptional repression domain (Figure 1B).[4,5] Leukemic *KMT2A* fusions impair H3K4 methylation and transform hematopoietic cells very efficiently.

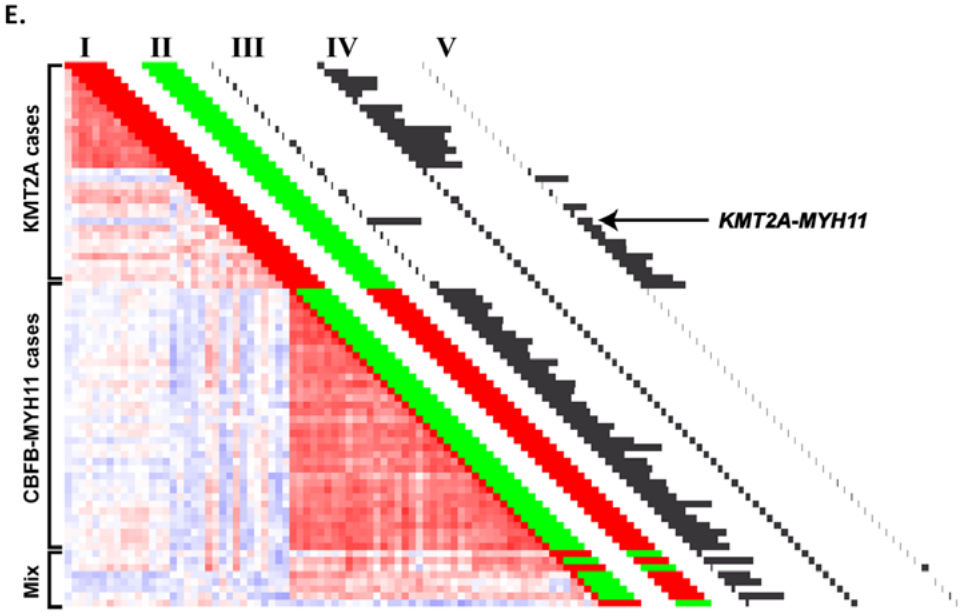
MYH11 is a subunit of a major contractile protein consisting of two heavy chain subunits and two pairs of non-identical light chain subunits. An pericentric inversion or translocation of chromosome 16 (inv(16)(p13q22) or t(16;16)(p13;q22)), involving the *MYH11* gene, define a specific subtype of AML characterized by eosinophilia and favorable treatment outcome and are characteristic for core-binding factor (CBF) leukemias.[1] These chromosome 16 abnormalities result in fusion of *MYH11* and core-binding factor β (*CBFB*) on 16q22. The resulting fusion transcript *CBFB-MYH11* encodes a protein consisting of the first 133-165 residues of the N-terminus of CBFB and variable C-terminal portions of MYH11. There are two models proposed for CBF-leukemogenesis both based on impairment of the master regulator RUNX1.[6] Briefly, the CBFB-MYH11 fusion protein affects RUNX1 either by sequestering RUNX1 from its target genes or interfering with RUNX1-mediated gene expression by binding of transcriptional repressors to the MYH11 moiety in CBFB-MYH11.[6]

The most frequent *CBFB-MYH11* fusions in adult AML fuses exon 5 of *CBFB* to exon 12 (type A,>85%), exon 8 (type D,<5%) or exon 7 (type E,<5%) of *MYH11*.[7] Several other fusions of *CBFB* and *MYH11* have been demonstrated, however, these are relatively rare (<1%). The variability among *CBFB-MYH11* fusions makes routine detection of this favorable marker in AML by RT-PCR challenging. By applying RT-qPCR aimed for expression of the 3'end of *MYH11,* which is normally not or at very low level expressed in hematopoietic cells, all AML inv(16) cases are reliably detectable, independent of the type of *CBFB-MYH11*.[8]

RESULTS

Here we describe a 67 year old patient who presented with MDS. After informed consent, bone marrow aspirates and peripheral blood samples were taken at diagnosis and at relapse. Cytological blood smear examination at diagnosis demonstrated a shift to the left in the blood smear with 7% myeloblasts, as confirmed by flow cytometry. Hypogranulated neutrophils were observed as were Pseudo Pelger-Huet nuclei and occasionally Auer rods. White blood cell count (WBC) was $13.5x10^9$/l. The bone marrow smears were hypercellular with 96% myelopoietic cells and 2% myeloblasts. Dysmyelopoiesis was seen, however, no increase of abnormal eosinophils.


A.
MDS
AML
B.
KMT2A
BCR
ATH 1-3
SNL 1-2
RD 1-2
PHD 1-3
TA
SET
KMT2A-MYH11
SH3 like
myosin head
myosin tail
MYH11
C.
marker
AML #5291
AML #5291
HL60
HEL
marker
500bp
400bp
300bp
200bp
D.
KMT2A exon 8
MYH11 exon 2
KMT2A [exon 8]
MYH11 [exon 2]
CACCACCAGAATCAG GGGACCAACAAGGCACCATGGCGCAGAAGG
P--P--P--E--S-- G--D--Q--Q--G--T--M--A--Q--K--


E.
I
II
III
IV
V
KMT2A cases
CBFB-MYH11 cases
Mix
KMT2A-MYH11

The patient was diagnosed as MDS-RAEB II according to the WHO 2008 classification. A diagnosis MDS was considered based on the elevated WBC and the shift to the left. The karyotype of the patient at diagnosis was 53,XY,+6,+8,+9,+13,+14,+19,+21[15]. Standard FISH both on interphase nuclei and metaphases using probes for *KMT2A*/11q23 (break apart), centromere 7 and 8, and probes for 5p15.2 and 5q31 revealed a translocation of *KMT2A*/11q23 to chromosome 16p13. The t(11;16)(q23;p13) was present in 96% of all cells (LSI MLL Dual color break apart probeset (Vysis)). The patient was treated according to the HOVON43 protocol (http://www.hovon.nl) and a complete remission was achieved, but the patient relapsed after 41 months. At relapse the patient was treated with AS602868, a pharmacological inhibitor of the IKK2 kinase, in a Phase 1 trial, but succumbed after progression of the AML. Cytological examination at relapse showed a hypercellular bone marrow, 72% myeloblasts, dysmyelopoiesis and dysmegakaryopoiesis (Figure 1A). Again Auer rods were seen but no eosinophilia or abnormal eosinophils. The karyotype was 53,XY,+6,+8,+9,+13,+14,+19,+21[1]/46,XY[19] FISH demonstrated the t(11;16)(q23;p13) to be present in 90% of all interphases.

All our AML cases are screened with RT-qPCR to detected possible *CBFB-MYH11* fusions. Interestingly, although the patient did not show an inv(16)(p13q22), t(16;16)(p13;q22) by cytogenetic analysis or *CBFB-MYH11* by RT-PCR, *MYH11* was highly expressed at diagnosis and at relapse shown by RT-qPCR[8], suggesting that *MYH11* was part of a unknown fusion transcript between 11q23 and 16p13. However, morphologically this case did not show any signs of inv(16)-associated eosinophilia.

To unravel the composition of the *MYH11*-containing mRNA transcript, we performed RNA sequencing (RNA-Seq) on patient material at diagnosis as part of our ongoing AML research. In brief, total sample RNA was extracted with phenol chloroform and reverse transcribed using Superscript II RT (Life Technologies). The cDNA was sheared with the Covaris device and further processed according to the TruSeq RNA Sample Preparation v2 Guide (Illumina). Amplified sample libraries were paired-end sequenced (2x36bp) on the HiSeq 2000 system and aligned against the human genome (hg19) using TopHat2.[9]

Figure 1. RNA-Seq reveals the *KMT2A-MYH11* fusion transcript (A) May-Grünwald-Giemsa staining of bone marrow from MDS and AML patient #5291. (B) Schematic representation of KMT2A, MYH11 and the KMT2A-MYH11 fusion protein. The KMT2A-MYH11 contains three short DNA-binding AT-hook motifs (ATH 1–3), two speckled nuclear localization sites (SNL1 and SNL2) and a transcriptional repression domain (TRD) followed by full length MYH11 (plant homology domain (PHD), transcriptional activation (TA) domain, methyltransferase domain (SET). The main *KMT2A* breakpoint region (BCR) is indicated.[5] (C) RT-PCR for the *KMT2A-MYH11* fusion transcript in (patient #5291 (duplicate)); cell lines HL60 and HEL (upper: primer set 13-561/562, lower: primer set 13-563/564) as negative controls. (D) Sanger sequencing of the *KMT2A* (exon 8) and *MYH11* (exon 2) fusion gene. (E) Pearson's Correlation View with pair-wise correlations between AML patients with KMT2A-rearranged AML [KMT2A cases], AML patients with inv(16) [*CBFB-MYH11* cases] and patient #5291 [*KMT2A-MYH11*] (indicated with arrow). The cells in the visualization are colored by Pearson's correlation coefficient values with deeper colors indicating higher positive (red) or negative (blue) correlations. Molecular data are depicted in the columns along the Correlation View: (I) *KMT2A* rearrangement and (II) *CBFB-MYH11* fusion (red bar: present and green bar: absent); gene expression of *MYH11* (III; 201497_x_at), *BRE* (IV; 205550_s_at) and *EVI1* (V; 221884_at). The bars are proportional to the level of expression.

All reads were aligned against genes annotated within the Ensembl database, as provided with the TopHat package. The Integrated Genome Viewer[10] was used for data visualization and assessment of *MYH11* fusion transcripts. The alignment of the RNA-Seq data confirmed the overexpression of *MYH11*, starting in exon 2 upstream of the ATG-start codon. Interestingly, paired-end sequencing reads of exon 2 of *MYH11* aligned to sequences of exon 8 of *KMT2A*. These results suggested that the t(11;16) resulted in a gene fusion of *KMT2A* (exon 8) and *MYH11* (exon 2) (Figure 1B). The *KMT2A-MYH11* fusions were confirmed by cDNA amplification using the primer sets 13-561MLL-MYH11 FW1 (*KMT2A* ex7): 5'-TTCCAGGAAGTCAAGCAAGC-3' and 13-562MLL-MYH11 RV1 (*MYH11* ex2): 5'-CTCGAAGCCCTGCTTCTC-3' (amplicon:298bp) or 13-563MLL-MYH11 FW2 (*KMT2A* ex7): 5'-CCGTCGAGGAAAAGAGTGAA-3' and 13-564MLL-MYH11 RV2 (*MYH11* ex2): 5'-CGTGACCTTCTTGCCATTCT-3' (amplicon:443bp) (0.25mM dNTP, 15pmol primers, 2mM $MgCl_2$, Taq polymerase and 1xbuffer [Life Technologies]). Cycling conditions: 1 cycle 5 min at 94˚C, 35 cycles 1 min at 94˚C, 1 min at 60˚C, 1 min at 72˚C, and 1 cycle 7 min at '72˚C. PCR amplification with both primer sets resulted in products with the expected size (Figure 1C). These PCR products were sequenced by using forward and reverse primers on the ABI PRISM3100 genetic analyzer (Applied Biosytems Life Technologies). Sanger sequencing confirmed the *KMT2A-MYH11* fusion transcript encoding an in-frame KMT2A-MYH11 fusion (Figure 1D). The *KMT2A-MYH11* fusion was demonstrated to be present at relapse (data not shown). Lack of high quality protein lysates prevented detection of the KMT2A-MYH11 fusion protein in the patient's MDS and AML phase.

Gene expression analyses demonstrated that the *KMT2A-MYH11* AML did not show any correlation with *CBFB-MYH11* AML (Figure 1E).[11,12] Interestingly, however, based on gene expression the *KMT2A-MYH11* AML grouped together with *MLL*-rearranged AML (Figure 1E). More specifically, the *KMT2A-MYH11* AML clustered with *MLL*-rearranged AML with high *EVI1* expression[13] instead of high *BRE* expression.[14]

Extensive analyses of the RNA-Seq data demonstrated a mutation in a well-known AML- and MDS-related gene, a non-synonymous mutation in the splicing factor gene *U2AF1* (exon2:c.C101T:p.S34F). This mutation in *U2AF1* has been confirmed by Sanger sequencing. Whether this mutation is somatic or germline remains to be elucidated.

DISCUSSION

In the past several t(11;16) patients have been described, however, these cases appear to be rare. In a study of two MDS patient with a t(11;16)(q23;p13) a recurrent fusion of the genes encoding KMT2A and CREB-binding protein (CREBBP (CBP)) was demonstrated.[15] A subsequent study of eight patients revealed that the t(11;16)(q23;p13) occurred exclusively in patients with therapy related t-AML or t-MDS, i.e., previous treatment with Topo2 inhibitors for a variety of malignancies.[16] Although the breakpoint was not determined in all t(11;16)(q23;p13) cases, it is unlikely that these patients carried a *KMT2A-MYH11* fusion considering the FISH probes used. Furthermore, our patient did not receive any treatment for any prior malignancy.

In this MDS/AML patient we have revealed a unique fusion of the N-terminal part of KMT2A and the complete *MYH11* protein. This fusion involves two proteins known to be present in chromosomal translocations in highly distinct AML subtypes. All *KMT2A* fusions are subdivided in 4 groups based on the *KMT2A*-fusion partner.[4] MYH11 contains several repeated helical rod domains important for self-dimerization and multimerization in its C-terminus, which also binds transcriptional corepressors. The novel KMT2A-MYH11 fusion most probably belongs to group 2 of MLL-fusion proteins, including SH3GL1/EEN, MMLT4/AF6, GAS7 and AFX1/FOXO4, which all contain oligomerization domains important for transformation. The presence of the t(11;16)(q23;p13) fusion at diagnosis and relapse suggests that KMT2A-MYH11 occurred in the founding clone of the MDS/AML and appears essential for this disease.

AUTORSHIP AND DISCLOSURES

MAS: Performed research, analyzed data and wrote paper; FGK: Performed research and analyzed data; AZ: Performed research and analyzed data; ASAAH: Performed research and analyzed data; SA: Performed research and analyzed data; HBB: Performed research and analyzed data; KvL: Performed research, analyzed data and wrote paper; PJMV: Designed and performed research, analyzed data and wrote paper. All authors declare no conflicts of interest

REFERENCES

1. Marcucci G, Haferlach T, Dohner H. Molecular Genetics of Adult Acute Myeloid Leukemia: Prognostic and Therapeutic Implications. *J Clin Oncol*. 2011;29(5):475-486.
2. Papaemmanuil E, Gerstung M, Malcovati L, et al. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*. 2013;122(22):3616-3627; quiz 3699.
3. Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.
4. Krivtsov AV, Armstrong SA. MLL translocations, histone modifications and leukaemia stem-cell development. *Nat Rev Cancer*. 2007;7(11):823-833.
5. Meyer C, Kowarz E, Hofmann J, et al. New insights to the MLL recombinome of acute leukemias. *Leukemia*. 2009;23(8):1490-1499.
6. Goyama S, Mulloy JC. Molecular pathogenesis of core binding factor leukemia: current knowledge and future prospects. *Int J Hematol*. 2011;94(2):126-133.
7. van Dongen JJ, Macintyre EA, Gabert JA, et al. Standardized RT-PCR analysis of fusion gene transcripts from chromosome aberrations in acute leukemia for detection of minimal residual disease. Report of the BIOMED-1 Concerted Action: investigation of minimal residual disease in acute leukemia. *Leukemia*. 1999;13(12):1901-1928.
8. van der Reijden BA, Massop M, Tonnissen E, et al. Rapid identification of CBFB-MYH11-positive acute myeloid leukemia (AML) cases by one single MYH11 real-time RT-PCR. *Blood*. 2003;101(12):5085-5086.
9. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
10. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.
11. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1617-1628.
12. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*. 2009;94(1):131-134.
13. Groschel S, Schlenk RF, Engelmann J, et al. Deregulated Expression of EVI1 Defines a Poor Prognostic Subset of MLL-Rearranged Acute Myeloid Leukemias: A Study of the German-Austrian Acute Myeloid Leukemia Study Group and the Dutch-Belgian-Swiss HOVON/SAKK Cooperative Group. *J Clin Oncol*. 2012.
14. Noordermeer SM, Sanders MA, Gilissen C, et al. High BRE expression predicts favorable outcome in adult acute myeloid leukemia, in particular among MLL-AF9-positive patients. *Blood*. 2011;118(20):5613-5621.
15. Taki T, Sako M, Tsuchida M, Hayashi Y. The t(11;16)(q23;p13) translocation in myelodysplastic syndrome fuses the MLL gene to the CBP gene. *Blood*. 1997;89(11):3945-3950.
16. Rowley JD, Reshmi S, Sobulo O, et al. All patients with the t(11;16)(q23;p13.3) that involves MLL and CBP have treatment-related hematologic disorders. *Blood*. 1997;90(2):535-541.

CHAPTER 10

Integrated genome-wide genotyping and gene expression profiling reveals *BCL11B* as a putative oncogene in acute myeloid leukemia with 14q32 aberrations

Saman Abbas[1], Mathijs A. Sanders[1], Annelieke Zeilemaker[1], Wendy M.C. Geertsma-Kleinekoort[1], Jasper E. Koenders[1], Francois G. Kavelaars[1], Zabiollah G.Abbas[1], Souad Mahamoud[1], Isabel W.T. Chu[1], Remco Hoogenboezem[1], Justine K. Peeters[1], Ellen van Drunen[2], Janneke van Galen[2], H. Berna Beverloo[2], Bob Löwenberg[1], and Peter J.M. Valk[1]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands
[2] Erasmus University Medical Center, Department of Clinical Genetics, Rotterdam, The Netherlands

Haematologica. 2014 May;99(5):848-57

ABSTRACT

Acute myeloid leukemia (AML) is a neoplasm characterized by recurrent molecular aberrations traditionally demonstrated by cytogenetic analyses. We have used high density genome-wide genotyping and gene expression profiling to reveal acquired cryptic abnormalities in AML. By genome-wide genotyping of 137 primary AML cases, we disclosed a recurrent focal amplification on chromosome 14q32, which included the *BCL11B, CCNK, C14orf177* and *SETD3* genes, in two cases. The *BCL11B* gene showed consistent high mRNA expression in the affected cases, whereas the expression of the other genes was unperturbed. Fluorescence in situ hybridization on 40 AML cases with high *BCL11B* mRNA expression (2.5-fold above median; 40 out of 530 cases (7.5%)) revealed 14q32 abnormalities in 2 additional cases. In the 4 *BCL11B*-rearranged cases the 14q32 locus was fused to different partner chromosomes. In fact, in 2 cases, we demonstrated that the focal 14q32 amplifications were integrated into transcriptionally active loci resulting in increased expression of full-length BCL11B protein. The *BCL11B*-rearranged AMLs expressed both myeloid and T-cell markers and all carried *FLT3* internal tandem duplications, a characteristic marker for AML. Generally, in AML, *BCL11B* mRNA expression appeared to be strongly associated with expression of other T-cell specific genes. Myeloid 32D(GCSF-R) cells ectopically expressing Bcl11b showed decreased proliferation rates and less maturation. In conclusion, by an integrated approach involving high-throughput genome-wide genotyping and gene expression profiling we identified *BCL11B* as a candidate oncogene in AML.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter10/

INTRODUCTION

AML is a heterogeneous clonal neoplasm characterized by accumulated genetic aberrations, which result in enhanced proliferation, a block in differentiation and increased survival of the leukemic blast cells and variable response to therapy.[1,2] In the past decades a number of recurrent cytogenetic abnormalities have been identified in AML, such as the chromosomal aberrations t(8;21) and inv(16).[1,2] These recurrent molecular lesions result in the expression of fusion proteins of which the leukemic potential, in combination with additional genetic events, has been demonstrated by *in vitro* and *in vivo* models.[3] In addition to cytogenetic abnormalities, acquired mutations in disease genes, such as *FLT3, NPM1, RUNX1* and *CEBPA,* have recently been demonstrated to be involved in AML.[1,2] Several acquired molecular aberrations carry prognostic value and have been incorporated in routine molecular analyses of AML.[1,2]

Contemporary genome-wide approaches, such as gene expression profiling (GEP), genome-wide genotyping and next generation sequencing (NGS), enable detailed analyses of hematologic malignancies for the identification of novel pathogenic genes.[2,4,5] For example, gene mutations in *IDH1*, *TET2*, *DNMT3A, ASXL1*, and *EZH2* have been demonstrated with these novel technologies in myeloproliferative neoplasms.[6-11] In addition to balanced translocations, loss or gains of genetic material are apparent in the leukemic blast of AML patients, e.g., those involving the partial or complete loss of chromosome 5 and 7.[1] In the past two decades, attempts to identify the tumor suppressor genes located on these chromosomes have failed. By genome-wide SNP genotyping it has become possible to simultaneously genotype hundred thousands of single nucleotide polymorphisms (SNPs) in a single assay. In addition, SNP platforms can also be conveniently used to determine chromosomal copy numbers, similarly to array comparative genomic hybridization (CGH). Genomic DNA can be examined with an inter-marker distance of several hundreds of base pairs, which makes it feasible to detect (micro)deletions and/or amplifications that are missed with conventional cytogenetics. The application of high-throughput SNP genotyping has been elegantly demonstrated to be powerful for the identification of disease genes, such as for ALL[4,12,13] or AML.[14] Another major advantage of SNP arrays is the fact that allele losses are directly recognizable as loss-of-heterozygosity (LOH). In fact, SNP arrays revealed that approximately 20% of AMLs exhibit large non-random regions of homozygosity without changes in copy number as a result of segmental uniparental disomy (UPD), often indicating mutations in genes within these regions. These areas of UPD have been associated with mutations in *CEBPA, WT1, FLT3* and *RUNX1*.[15,16] In addition, deletions, amplifications and UPDs could alter the gene expression levels of proximal genes. Juxtaposition of regulatory sequences may result in increased or decreased expression of affected genes. Thus, genome-wide analyses to detect copy number changes and LOH in the context of gene expression may also pinpoint towards pathogenic genes. We recently developed SNPExpress, an easily accessible software tool to accurately analyze SNP genotype calls, copy number and gene expression in an efficient combinatorial way.[17]

In this study, we identified *BCL11B* as a candidate oncogene in AML through an integrated approach of genome-wide genotyping and GEP, followed by NGS. *BCL11B* is a Kruppel family zinc finger family gene located at 14q32, associated with transcriptional co-repressor complexes in mammalian cells and a pivotal regulator of differentiation and survival of haematopoietic cells, especially T-cells.[18] We demonstrate that *BCL11B* is involved in a number of cryptic 14q32 translocations in AML, in which *BCL11B* and T-cell associated genes expression levels are increased concomitantly. Overexpression of BCL11B in a murine myeloid cell line model inhibits proliferation.

METHODS

Patients samples

This study has been approved by our local Medical Ethical Committee (MEC-2004-030 and MEC2007-364). After informed consent, bone marrow aspirates or peripheral blood samples of a representative cohort of AML patients were collected. Eligible patients had a diagnosis of primary AML, confirmed by cytological examination of blood and bone marrow. All patients were treated according to the HOVON (Dutch-Belgian Hematology-Oncology Co-operative group) protocols (http://www.hovon.nl). For details see Supplementary material.

Genome-wide genotyping and gene expression profiling

Genome-wide genotyping data sets of 48 patients with various subtypes of AML were generated using Affymetrix 500K *Nsp*I/*Sty*I DNA mapping arrays and 89 patients with cytogenetically normal AML (CN-AML) using Affymetrix 250K *Nsp*I or *Sty*I DNA Mapping arrays. The copy numbers of all AML samples were calculated using diploid references, i.e., 15 normal karyotype AML samples. For details see Supplementary material.

Gene expression profiles of the same AMLs were generated using Affymetrix HG-U133 plus 2.0, as described elsewhere (GEO Series accession number GSE6891).[19] Pearson correlation analyses was performed as described previously.[20] The genome-wide genotyping and gene expression profiling data sets were examined using SNPExpress.[17]

Fluorescence In Situ Hybridization (FISH)

Dual color fluorescence *in situ* hybridization (FISH) was performed with BAC clones RP11-431B1, RP11-876E22, RP11-830F3, RP11-782I5, RP11-450C22, RP11-57E12, RP11-1069L3 and RP11-242A7 covering the *BCL11B* encompassing region and regions up- and downstream (BACPAC resources, Oakland, USA). For details see Supplementary material.

Targeted sequencing of the 14q32 genomic region

Library preparation and targeted resequencing was performed following the protocols as described previously.[21] In brief, high molecular weight DNA of AML #2301 and #7073 were sheared using a Covaris E210 waterbath sonificator. The *BCL11B* 14q32 – tel. genomic region (chr14:93930247-105928955 (hg19)) was captured with a Roche/Nimblegen SeqCap EZ Choice XL Library. The captured region was subsequently paired-end sequenced using the Illumina HiSeq2000. The data has been analyzed using an in-house pipeline which identifies single nucleotide variants, small and large indels and copy number variations. The chromosomal breakpoints, in the 14q32 region and the partner chromosome, were identified using Breakdancer.[22] The genomic fusions were subsequently confirmed by Sanger sequencing.

Western blot analyses

Western blot analyses were carried out using an affinity-purified rabbit polyclonal anti-BCL11b antibody (Novus Biologicals, Littleton, USA). Immune complexes were detected by binding anti-mouse IgG conjugated to horseradish peroxidase (DAKO, Heverlee, Belgium) followed by the enhanced chemiluminescence assay (Amersham Bioscience, Piscataway, NJ) and GAPDH was stained with primary affinity-purified rabbit polyclonal antibody (α-GAPDH FL-335) (Santa Cruz Biotechnology, California, USA). For details see Supplementary material.

DNA constructs and generation of BCL11B expressing 32D/GCSFR cells

Murine *Bcl11b* cDNA (kindly donated by Dorina Avram, Albany Medical Center, Albany, NY) was subcloned into a pLXSN expression vector under control of a 5' long terminal repeat (LTR) of the Moloney murine sarcoma virus (MoMSV) (Clontech, Mountainview, USA). Vector constructs were confirmed by nucleotide sequencing and retrovirally transfected into 32D cells that stably express human granulocyte colony-stimulating factor receptor (GCSF-R)[23] using Fugene transfection reagent (Roche, Indianapolis, USA). Cells were stimulated with interleukin-3 (IL3, 25ng/ml) or GCSF (25 ng/ml), counted and assessed for proliferation and granulocytic differentiation. For details see Supplementary material.

RESULTS

Genome-wide genotyping of cytogenetically abnormal and normal AML cases

In total, DNA mapping array profiles of 137 AML cases were generated (Figure 1), containing 48 AML cases selected based on previous GEP studies, i.e., 21 AML cases from GEP clusters #4 and #15 (100% *CEBPA* mutant or *CEBPA* silenced[24]), 13 AML cases from GEP cluster #9 (100% inv(16)) and 14 AML cases from GEP cluster #10 (adverse prognosis).[20] In addition, DNA mapping array profiles,

i.e., Affymetrix 250K *Nsp*I or *Sty*I DNA mapping array, of 89 CN-AML cases were generated. With the Affymetrix 500K *Nsp*I/StyI DNA Mapping arrays, all known numerical cytogenetic aberrations, i.e., whole chromosome and interstitial deletions and amplifications that had been identified with cytogenetic banding analysis, were recognized in the 48 cytogenetically abnormal AML samples, as long as the abnormalities were present in over 30% of the AML cells.[17] Also, in approximately 25% of all cases large regions of segmental uniparental disomy were detected, often involving whole chromosome arms.[17]

In addition to the known cytogenetic aberrations, relatively low numbers of small interstitial deletions and amplifications were detected in the 137 AML cases. However, some of these were indicative for the presence of cryptic translocations, such as cryptic t(5;11), t(9;22) and t(4;11), which are known to encode chimeric fusion proteins essential for leukemogenesis. All fusion transcripts involved in these translocations, i.e., *NUP98-NSD1*[25], *BCR-ABL* and *MLL-AF4*, were confirmed by RT-PCR. Thus, although relatively small numbers of aberrations were found, most being non-recurrent, they may reliably mark relevant leukemic lesions.

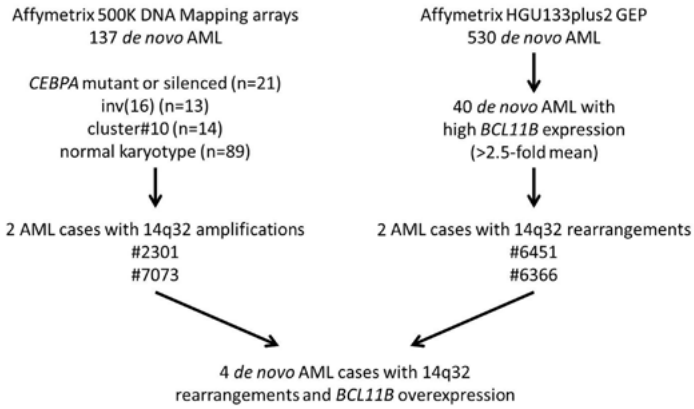


Figure 1. Research design. GEP: gene expression profiling.

Integrated analysis of genome-wide genotypes and gene expression profiles

By an integrated approach using genome-wide genotyping data and previously determined GEPs of the primary AML samples (Figure 1)[20], we searched for genes aberrantly expressed as a result of numerical changes in the AML genome. Using SNPExpress[17], we identified 2 AML cases with relatively small interstitial amplifications in the 14q32 region (#2301 amplification: 482 Kb, 3 copies and #7073 amplification: 460 Kb, 3 copies) (Figures 2A and B). The amplified region encompassed *BCL11B, CCNK, C14orf177* and *SETD3*. Interestingly, *BCL11B* mRNA was highly expressed in the 2 AML cases with numerical changes, whereas expression of *C14orf177*, *CCNK*

and *SETD3* were unperturbed compared to other AML cases (Figures 2A and B). In addition, *BCL11B* mRNA is highly expressed in AML #2301 and #7073, whereas expression is low or absent in control AML cases (Figure 2A). This could indicate that as a result of a genomic rearrangement, *BCL11B* has become overexpressed in these AML cases. The small interstitial amplifications in AML #2301 and #7073 may pinpoint towards cryptic translocations.

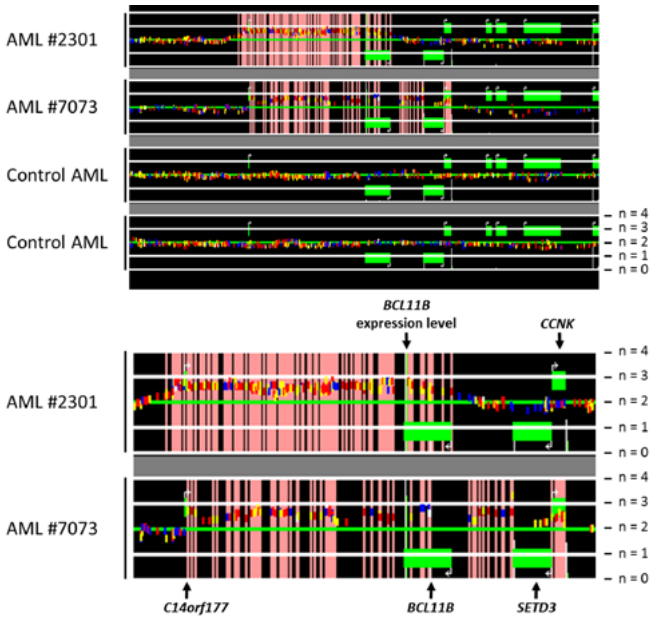


Figure 2. Identification of interstitial amplifications on 14q32.2 using SNPExpress. (A) Copy number profiles of chromosome 14q32.2 for four AML samples. Copy numbers are shown for each patient by horizontal lines (n=0, 1, 2, 3, 4). SNP genotypes are color coded: AA: red; BB: yellow; AB: blue, No call: white. Gains (default n>2.5) are depicted as pink background. Gene expression levels are visualized as vertical white bars. Multiple probe sets spanning the same locus is depicted by a green bar proportional to the highest expression value observed. Green boxes represent genes and accompanying arrow indicates its orientation. In AML #2301 and AML #7073 clear amplifications are visible, whereas these aberrations are absent in the two control AMLs. (B) Snapshot of SNPExpress showing the amplified region in AML case #2301 and #7073 including genes located within this region. *C14orf177* and *BCL11B* are amplified in both AML cases, whereas *SETD3* and *CCNK* only in AML #7073. *BCL11B* expression is increased in AML #2301 and #7073, as indicated by the green bar (multiple probe sets), and absent in control AML cases (Figure 2A).

FISH reveals translocations in AML cases #2301 and #7073 involving *BCL11B*

To confirm the amplifications in the *BCL11B* locus in the 2 AML cases, we performed FISH analysis with a probe covering the *BCL11B* gene (RP11-431B1) and a probe flanking this locus (RP11-74H1) (Figure 3A). On metaphase spreads of both AML cases an additional *BCL11B* allele was apparent (Figure 3B). This is in line with the expected copy number change for the *BCL11B* locus (n=3) as

shown with SNPExpress (Figures 2A and B). In fact, through verification using chromosomal paints we showed that *BCL11B* was translocated to chromosome 6 in AML case #2301 and chromosome 8 in AML #7073 (data not shown).

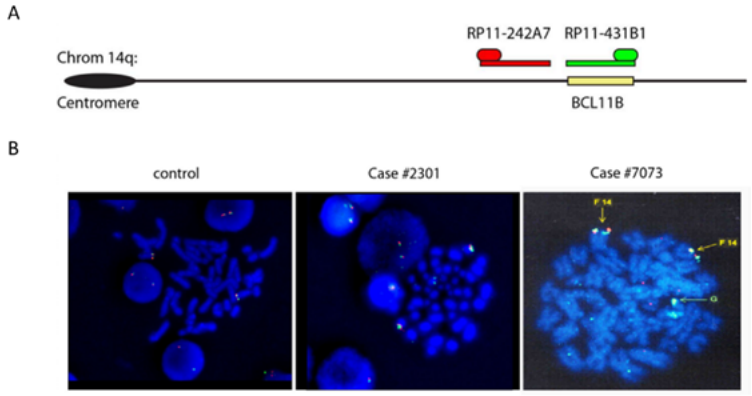


Figure 3. FISH analysis of AML cases #2301 and #7073 using probes specific for *BCL11B* and its flanking region. (A) Schematic representation of the FITC-labeled BAC probe (RP11 431B1) covering the *BCL11B* locus and Texas Red-labeled BAC probe (RP11 242A7) covering the adjacent region. (B) Microscope images of FISH analysis performed on metaphases chromosomes of AML cases #2301 and #7073 showing additional green signal (RP11431B1) indicative for an extra copy of the *BCL11B* locus.

Amplified 14q32 genomic regions are integrated into transcriptionally active loci

We performed paired-end sequencing of the 14q32 captured region in AML cases #2301 and #7073. We observed paired-end reads spanning 14q32 and 6q25.3 (chr6:156717480 and chr14:99110325; chr6:156587275 and chr14:99748893) in AML case #2301 and reads spanning 14q32 and 8q24.21 (chr8:130485869 and chr14:99179210) in AML case #7073, indicating translocation events to partner chromosomes 6 and 8, respectively. These breakpoints were confirmed by PCR on genomic DNA, followed by Sanger sequencing. The *BCL11B* encoding-amplified DNA integrated into two transcriptionally active regions on chromosome 6 and 8, i.e., on 6q25.3 into an expressed sequence tag *CB984582* and on 8q24.21 into the large non-coding (lnc) RNA gene *Coiled-Coil Domain Containing 26* (*CCDC26*). Both polyA genes are transcriptionally active in the 2 AML cases, and various other AML cases, and are subjected to mRNA splicing as demonstrated by RNA sequencing (data not shown), indicating that these RNAs are expressed in myeloid cells. No fusion transcripts between *BCL11B* and RNAs encoded by the partner chromosomes could be detected by RNA-seq, suggesting that regulatory sequences on 6q25.3 and 8q24.21 may activate the *BCL11B* gene in the *BCL11B*-rearranged AML cases.

AML case #2301 expresses full-length BCL11B

The translocations involving *BCL11B* could result in increased expression of either full-length BCL11B or a fusion protein involving BCL11B. In fact, *BCL11B* mRNA expression in AML#2301 and AML#7073, were, respectively, 12- and 8-fold over mean *BCL11* mRNA expression in 530 AML cases (219528_s_at; 22895_s_at; 224310_s_at).[19] Next, we examined the expression profiles obtained with Affymetrix Human Exon 1.0 ST Array for AML case #2301. This analysis showed that in AML #2301 all four exons of *BCL11B* were expressed at similarly high levels (data not shown). The fact that exon 1 of *BCL11B,* containing the ATG start codon was expressed, suggested that full-length BCL11B is expressed rather than a fusion protein involving parts of BCL11B. Western blot analyses of whole, cytoplasmic and nuclear cell lysates of AML case #2301 were used to assess both the size and localization of the BCL11B protein. Immunodetection with BCL11B antibodies confirmed the expression of full length BCL11B protein restricted to the nucleus (Figure 4). Of note, full-length BCL11B was also highly expressed in AML case #2238, an AML without any known aberration involving *BCL11B*.

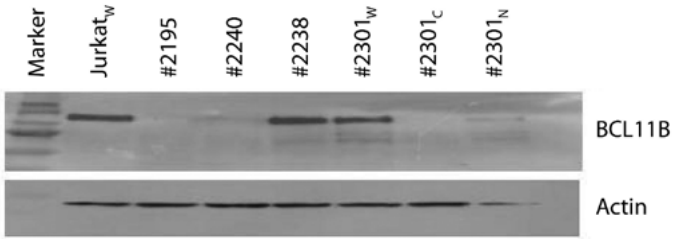


Figure 4. Western blot analysis of BCL11B in AML case #2301. Western blot analysis with a BCL11B-specific antibody demonstrates high expression of full length BCL11B in AML case #2301, in the nuclear compartment (upper panel). Whole cell lysates from Jurkat, an acute T-cell leukemia cell line, and AML #2238 show high BCL11B expression. AML cases #2195 and #2240 with low *BCL11B* mRNA expression were used as negative controls (#2301$_w$:whole cell lysate; #2301$_c$: cytoplasmic lysate; #2301$_n$: nuclear lysate). β-actin was used as loading control (lower panel).

FISH analyses of selected AML cases with high *BCL11B* mRNA expression reveals additional cases with *BCL11B* translocations

FISH analysis of AML cases with high *BCL11B* expression revealed translocations involving *BCL11B*, thereby raising the possibility that other AML cases with aberrantly high *BCL11B* expression would harbour *BCL11B*-rearrangements as well. GEP of 530 AML cases[19] showed variable expression of *BCL11B* mRNA in AML subsets, including case #2301 and #7073 (Figure 5A). We selected 40 AML cases with increased *BCL11B* mRNA expression, i.e., >2.5-fold above mean *BCL11B* expression in primary AML (Figure 5A), and performed FISH analysis on the *BCL11B* chromosomal region. FISH analyses revealed 2 additional AML cases with a *BCL11B* translocation (AML #6366 and #6451 (Figure 5A and B)). With specific chromosomal paints, we showed that in AML case #6451

the *BCL11B* locus was translocated to chromosome 7 (data not shown). Further FISH could not be carried out on AML #6366 due to the lack of material.

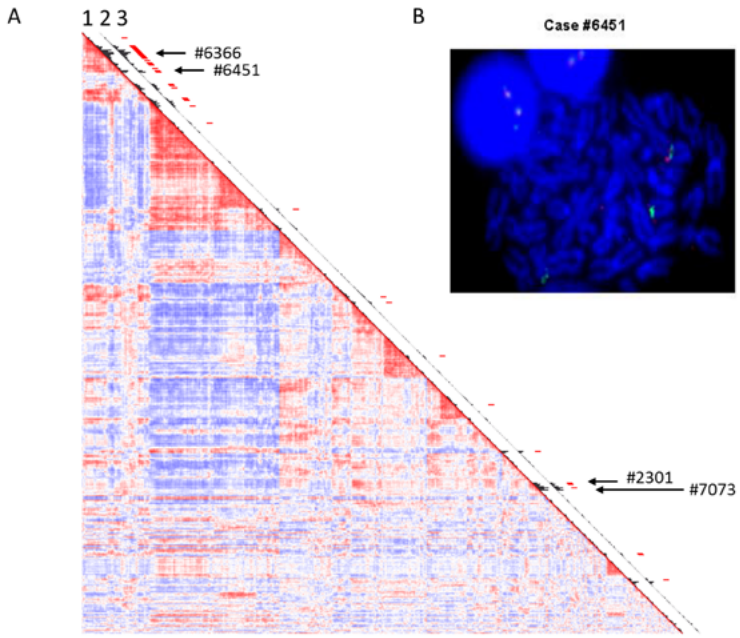


Figure 5. Correlation view based on gene expression profiling of 530 AML cases. (A) Pearson correlation view of 530 AML cases showing gene expression correlation based on 2847 probe sets. The black bars indicate expression of *BCL11B* 1: *BCL11B* expression: 219528_s_at and 2: *BCL11B* expression: 222895_s_at, where the size of the bars is proportional to the levels of *BCL11B* expression; 3: selection of AML with *BCL11B* overexpression (>2.5-fold mean). The *BCL11B*-rearranged cases #2301, #6451, #6366 and #7073 are indicated by an arrow. (B) FISH analysis performed on metaphase spreads of AML cases #6451 showing disassociation of the probe RP11242A7 (red) and RP11431B1 (green) indicating translocation of *BCL11B*.

Immunophenotyping and molecular analyses of AML cases carrying *BCL11B* aberrations

Immunophenotyping of the AML cases harbouring *BCL11B* translocations expressed, besides myeloid markers, also lymphoid markers such as CD2, CD3, and CD7 (Table 1). Cytoplasmic(cy) CD3 expression was present in case #2301, suggesting its classification as T-ALL, however, cyCD3 was absent in the remaining cases. In fact, these AML cases appeared to have a biphenotypic signature, i.e., expressing (early) myeloid as well as T-cell associated markers. Well-known recurrent molecular abnormalities were determined in the *BCL11B*-rearranged AML cases, demonstrating, with no exception, that these cases carried mutations in the *FLT3* gene (Table 1), i.e., internal tandem duplications (ITD) or mutations in the tyrosine kinase domain (TKD). We did not identify mutations in *K-RAS, N-RAS, c-KIT, IDH1, IDH2, ASXL1* or *CEBPA*. Case #6366 also carried a *DNMT3A* mutation (Table 1).

We have analyzed the immunophenotype of the *BCL11B* non-rearranged cases with *BCL11B* overexpression and did not find a specific pattern of T-cell specific markers. Several cases do express CD7, however, this aberrant marker is relatively frequent present on myeloid leukemic blasts (app. 30%). Moreover, we were unable to demonstrate a significant association between *BCL11B* overexpression and *FLT3*-ITD or -TKD mutations in *BCL11B* non-rearranged AML cases.

Table 1. Clinical, molecular and immunophenotypic data of the AML cases with *BCL11B* translocations.

Patient number	AML #2301	AML #7073	AML #6451	AML #6366
FLT3-ITD	pos	pos	pos	pos
FLT3 TKD835	neg	neg	pos	neg
DNMT3A mutation	neg	neg	neg	pos
FAB	M1	M4	M1	M2
WHO	1 WHO	1 WHO	2 WHO	0 WHO
Gender	M	M	F	F
Karyotype	46,XY[21]/ ?46,XY,inc[9]	46,XY[20]	46,XX,del(7)(q21q35) [5]/46,idem,add(13) (q3?4)[17]/ 46,idem,add(9)(q3?4) [2]/46,XX[15]	53,XX,+4,+8,+10,+13, +14,+15,+20[4]/ 46,XX[35]
Immunophenotype	CD45(+), HLA-DR-, CD34+, TdT+, MPOpartial+, CD1-,CD2+, CD3+,CD4-,CD5-	CD15partial+, CD33+,CD7+, CD36partial+, CD56-,CD65s-, CD117partial+, CD133+, CD4partial+	CD45(+), HLA-DR+,CD34+, TdTpartial+, MPOpartial+, CD11c-, CD13+,CD15-, CD15s partial+, CD33-, CD65s-, CD117+, CD133+, CD2+, CD7partial+	CD45(+),HLA-DR+, CD34partial+,TdT-, MPOpartial+, CD11c partial+, CD13partial+, CD15partial+, CD33+,CD36partial+, CD117 partial+, CD133+, CD4 partial+, CD7(+)

The mutation status for *FLT3*, *NPM1*, *N-RAS*, *K-RAS*, *CEBPA*, *c-KIT*, *ASXL1*, *IDH1*, *IDH2 and DNMT3A* was determined as described previously.[26-28] No mutations were present in the 4 AML cases in *NPM1, N-RAS, K-RAS, CEBPA, KIT, ASXL1*, *IDH1,* and *IDH2* (Pos: mutant; Neg: wild-type).

***BCL11B* is aberrantly expressed in AML and associated with T-cell gene expression signature**

To investigate whether other AML cases with elevated *BCL11B* mRNA expression show full-length BCL11B expression, we carried out Western blot analyses on a limited number of cases. All analyzed samples with high *BCL11B* mRNA showed full length BCL11B protein expression at variable levels (Figure 6). Due to lack of specimens, BCL11B protein expression analyses in non-rearranged *BCL11B* cases was limited to those shown in Figures 4 and 6. To examine which genes are co-expressed with *BCL11B* in AML, we performed a Pearson correlation analysis using GEP data of 530 AML cases.[19] *BCL11B* co-regulated probe sets were calculated across all AML patients. The top 50 *BCL11B* correlating probe sets are highly associated with T-cells and T-cell development (Supplementary table 1). In fact, the majority of *BCL11B* associated genes are T-cell specific genes, such as *CD3, TRBV19, IL32, LCK, TCF7* and *CD2*, among many others (Supplementary table 1).

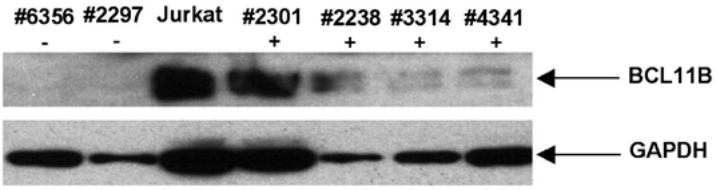


Figure 6. Western blot analyses for BCL11B primary AML. Immuno-detection of the BCL11B protein in AML cases with elevated levels of *BCL11B* mRNA (+) and cases with undetectable levels of *BCL11B* mRNA (-) (upper panel; Jurkat cell lysate as positive control). GAPDH was used as loading control (lower panel).

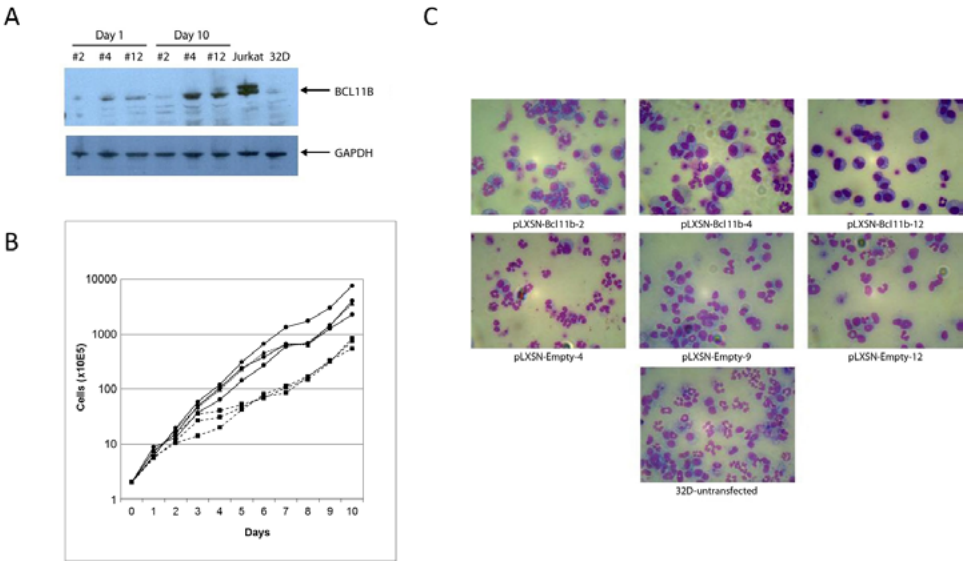


Figure 7. Effects of BCL11B overexpression in murine 32D(GCSF-R) cells. (A) Western blot analyses for BCL11B in 32D(GCSFR) cells. 32D(GCSFR) clones overexpressing BCL11B are indicated by #2, #4, and #12 (IL3 for 1 and 10 days). Lysates obtained from these clones were immunostained for BCL11B at day 1 and day 10 (Jurkat cells: positive control; 32D: 32D(GCSF-R) cells). GAPDH was used as loading control (lower panel). (B) Growth curve of 32D(GCSFR) cells with (squares, dashed line) and without (round, solid line) BCL11B expression and parental 32D(GCSFR) cells (triangle, dotted line) incubated with interleukin 3. 32D cells were counted every 24 h for ten days. (C) May-Grünwald-Giemsa-stained cytospins of 32D(GCSF-R) cells with (upper panel) and without (lower panel) BCL11B expression incubated with GCSF for 7 days. Granulocytic differentiation is monitored by the presence of cells with segmented nuclei.

Increased *Bcl11b* expression results in decreased proliferation of the myeloid cell line 32D(GCSF-R)

To investigate the effect of *Bcl11b* expression on proliferation and differentiation, immortalized myeloblast-like murine bone marrow cells stably expressing human GCSF-R (32D(GCSF-R)) were transfected with full length murine *Bcl11b* cDNA. Three 32D(GCSF-R) clones expressing *Bcl11b* were selected and incubated for ten days in the presence of interleukin-3 (IL3) or

granulocyte stimulating factor (GCSF). Western blot analyses demonstrated that BCL11B was expressed at every time point (Figure 7A). BCL11B expressing 32D(GCSF-R) clones showed a consistent decreased proliferation rate when cultured in the presence of IL3 in comparison to 32D(GCSF-R) clones containing a control empty vector (Figure 7B). Additionally, we evaluated the granulocytic differentiation abilities of the same 32D(GCSF-R) clones upon stimulation with GCSF. Morphological analyses of cytospins did not show consistent maturation defects in the BCL11B expressing 32D(GCSF-R) clones. However, less maturation towards granulocytes in 32D(GCSF-R) cells expressing BCL11B and more undifferentiated blast cells compared to cells with empty vector were present in some 32D(GCSF-R) BCL11B expressing clones (Figure 7C). This effect was most apparent in 32D(GCSF-R) BCL11B clones #4 and #12, the 32D(GCSF-R) clones with the highest expression levels of BCL11B (Figure 6A).

DISCUSSION

Integrative analyses of genome-wide genotyping and copy number data with GEP enables the identification of pathogenic genes aberrantly expressed due to genomic imbalances. The outlined integrative approach resulted in the identification of *BCL11B* as a novel oncogene in AML. Interstitial amplification of 14q32 was initially revealed in 2 AML cases by genome-wide genotyping encompassing the *BCL11B* gene. Although other genes were affected by these copy number changes, these 14q32 aberrations resulted in unique *BCL11B* mRNA and full-length BCL11B protein overexpression. FISH in a selection of 40 AML cases with high *BCL11B* mRNA expression identified 2 additional AML cases bearing *BCL11B* translocations.

The BCL11 family has two members BCL11A and BCL11B.[18] Bcl11a was identified as a common retroviral insertion site (Evi9) in murine myeloid leukemias and is required for normal B-cell development.[29] Mice carrying biallelic inactivation of *Bcl11b* developed thymic lymphomas, indicating that loss-of-function mutations in *Bcl11b* contribute to mouse lymphomagenesis and possibly to human cancer development.[30] *BCL11B* is a four-exon gene located on 14q32, encoding a Kruppel family zinc finger transcription factor and a key regulator of differentiation and survival of thymocytes.[18] *BCL11B* was first associated with hematological malignancies due to its recurrent involvement with the homeobox transcription factor *TLX3* in a relatively high percentage of pediatric and adult T-cell acute lymphoblastic leukemia (T-ALL) carrying the cryptic t(5;14)(q35;q32).[31,32] Less frequent, T-ALL samples with an inv(14)(q11.2q32.31) carry an in-frame transcript of *BCL11B* and the T-cell receptor gene segment *TRDV1*. These ALL cases do not express wild type *BCL11B* transcripts, suggesting that *BCL11B* disruption may contribute to T-cell malignancies in humans.[33] Interestingly, recently a DNA copy number and sequencing analyses approach revealed mono-allelic *BCL11B* deletions and missense mutations in 10-15% of T-ALL.[34,35] Structural homology modeling revealed that several of the BCL11B mutations disrupted the structure of the zinc finger domains required for DNA binding.

A number of myeloid, mixed-lineage, and non-lymphocytic leukemias with 14q32 abnormalities have been reported, however, in these instances the affected genes were not identified.[36-40] The first evidence of *BCL11B* involvement in 14q32 translocations in AML was reported by Bezrookove et al.[37] They reported one case of t(6;14)(q25~q26;q32) in an adult with AML and used bacterial artificial chromosomes to demonstrate the involvement of *BCL11B* in this AML case.[37] Due to lack of patient material, the investigators could neither establish the deregulation of *BCL11B* nor the identification of the partner genomic locus.[37] Of note, the breakpoint in this AML case appeared to be located upstream of the *BCL11B* gene. This is similar to the AML cases described here and suggests that the breakpoints in AML are clustered upstream, whereas in ALL they are downstream of *BCL11B*. Specific chromosomal paints demonstrated that different partner chromosomes were involved in the AML cases with a *BCL11B* translocation. The fact that in 2 AML cases the *BCL11B*-containing amplified region integrated in transcriptionally active lncRNAs may suggest that different regulatory regions of the lncRNAs are capable to activate the *BCL11B* oncogene. The breakpoints in AML #2301 and #7073 are 10kb and 600kb away from the transcriptional start site of *BCL11B*. There is no obvious reason why specifically *BCL11B* and not the other genes would become activated. Interestingly, however, the rearranged *BCL11B* allele in both AML #2301 and AML #7073 is juxtaposed to recently described super-enhancers, which have been shown to act as key oncogenic drivers.[41,42] These putative super-enhancers are present in the cell line MOLM-1 and seem to be linked to *ARID1B* on chromosome 6 (#2301) and *MYC/GSDMC* on chromosome 8 (#7073) and may be responsible for increased *BCL11B* expression. The selective overexpression of BCL11B may give the cells a specific advantage, whereas the other genes located on 14q32 would not. Interestingly, BCL11B protein appeared to be expressed in additional primary AML cases that did not carry *BCL11B* translocations. In these AML cases other mutations may be present or *BCL11B* may be activated by other means.

The 14q32 region, including *BCL11B*, has been subjected to translocations in T-ALL and acute mixed lineage leukemia.[23,36-40] In fact, the involvement of 14q32 translocations and *BCL11B* in AML has been debated.[43] However, the immunophenotyping and molecular analyses of the AML samples with *BCL11B* translocations described here showed that these leukemias have a biphenotypic immunophenotype, but also all carry a common AML-associated *FLT3*-ITD mutation. These leukemias do, therefore, share a characteristic genetic feature with AML.

BCL11B is expressed in T-lymphocytes and T-cell leukemias and is a pivotal regulator of a number of genes related to T-cell proliferation and differentiation such as *IL2*, *NF-kappaB, TCRβ* and *p21*.[44-48] It was shown recently that the expression of *BCL11B* in T-cell lines resulted in markedly increased apoptosis resistance following treatment with radiomimetic drugs accompanied by a cell cycle delay caused by accumulation of cells at G1.[49] We examined the consequences of *Bcl11b* overexpression on proliferation and differentiation in a mouse myeloid 32D(GCSF-R) cell line model. 32D(GCSF-R) cells, expressing full length murine *Bcl11b* cDNA, showed a consistent decreased proliferation rate compared to cells expressing the empty vector or to the parental

untransfected cells. Upon stimulation with GCSF, 32D(GCSF-R) cells overexpressing BCL11B showed less maturation towards granulocytes compared to cells expressing empty vector, giving supporting evidence that BCL11B is partially blocking or delaying differentiation in 32D(GCSF-R) cells. The decreased proliferation rate in BCL11B expressing cells may suggest that a proliferative mutation, such as a *FLT3*-ITD, may indeed be required for full leukemic transformation.

In conclusion, we show that *BCL11B* is involved in 14q32 translocations with different putative chromosomal partners in well-characterized AML cases using high-throughput genome-wide genotyping, cytogenetics and GEP. In these translocations, full length BCL11B is highly expressed concomitantly with T-cell specific markers. We speculate that due to the translocations, *BCL11B* expression is influenced by active transcriptional elements on the partner chromosomes resulting in high *BCL11B* expression and consequently T-cell associated genes. The murine cell line 32D(GCSF-R) overexpressing BCL11B shows decreased proliferation and partial delayed differentiation, which provides evidence that *BCL11B* may have suppressive and disruptive effects on cell proliferation and differentiation of myeloid cells. Altogether, these analyses revealed *BCL11B* as a putative oncogene in AML with and possibly without aberrations involving 14q32.

ACKNOWLEDGEMENTS

This work was supported by grants from the Dutch Cancer Society (Koningin Wilhelmina Fonds) and performed within the framework of CTMM, the Center for Translational Molecular Medicine (Leukemia BioCHIP project (grant 030-102)). We are indebted to the participants of the HOVON clinical trials, our colleagues from the stem cell transplantation and molecular diagnostics laboratories who provided, collected and analyzed AML cell samples.

AUTHORSHIP AND DISCLOSURES

SA: Performed research, analyzed data and wrote manuscript; MS: Analyzed data; AZ: Performed research; WMCG: Performed research; JEK: Performed research; FGK: Performed research; ZGA: Performed research; SM: Performed research; IWTC: Performed research; RH: Analyzed data; JKP: Analyzed data; EvD: Performed research; JvG: Performed research; HBB: Performed research and wrote manuscript; BL: Designed research and wrote manuscript; PJMV: Designed and performed research, analyzed data and wrote manuscript.

REFERENCES

1. Burnett A, Wetzler M, Lowenberg B. Therapeutic Advances in Acute Myeloid Leukemia. *J Clin Oncol*. 2011;29(5):487-494.
2. Marcucci G, Haferlach T, Dohner H. Molecular Genetics of Adult Acute Myeloid Leukemia: Prognostic and Therapeutic Implications. *J Clin Oncol*. 2011;29(5):475-486.
3. Goyama S, Mulloy JC. Molecular pathogenesis of core binding factor leukemia: current knowledge and future prospects. *Int J Hematol*. 2011.
4. Mullighan CG, Goorha S, Radtke I, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007;446(7137):758-764.
5. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008;456(7218):66-72.
6. Delhommeau F, Dupont S, Della Valle V, et al. Mutation in TET2 in myeloid cancers. *N Engl J Med*. 2009;360(22):2289-2301.
7. Langemeijer SM, Kuiper RP, Berends M, et al. Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat Genet*. 2009;41(7):838-842.
8. Gelsi-Boyer V, Trouplin V, Adelaide J, et al. Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br J Haematol*. 2009;145(6):788-800.
9. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med*. 2009;361(11):1058-1066.
10. Ernst T, Chase AJ, Score J, et al. Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat Genet*. 2010;42(8):722-726.
11. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med*. 2010;363(25):2424-2433.
12. Mullighan CG, Flotho C, Downing JR. Genomic assessment of pediatric acute leukemia. *Cancer J*. 2005;11(4):268-282.
13. Mullighan CG, Su X, Zhang J, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N Engl J Med*. 2009;360(5):470-480.
14. Bullinger L, Frohling S. Array-based cytogenetic approaches in acute myeloid leukemia: clinical impact and biological insights. *Semin Oncol*. 2012;39(1):37-46.
15. Raghavan M, Lillington DM, Skoulakis S, et al. Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. *Cancer Res*. 2005;65(2):375-378.
16. Gorletta TA, Gasparini P, D'Elios MM, Trubia M, Pelicci PG, Di Fiore PP. Frequent loss of heterozygosity without loss of genetic material in acute myeloid leukemia with a normal karyotype. *Genes Chromosomes Cancer*. 2005;44(3):334-337.
17. Sanders MA, Verhaak RG, Geertsma-Kleinekoort WM, et al. SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels. *BMC Genomics*. 2008;9:41.
18. Liu P, Li P, Burke S. Critical roles of Bcl11b in T-cell development and maintenance of T-cell identity. *Immunol Rev*. 2010;238(1):138-149.
19. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*. 2009;94(1):131-134.
20. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1617-1628.
21. Beekman R, Valkhof MG, Sanders MA, et al. Sequential gain of mutations in severe congenital neutropenia progressing to acute myeloid leukemia. *Blood*. 2012;119(22):5071-5077.
22. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6(9):677-681.

23. Dong F, van Buitenen C, Pouwels K, Hoefsloot LH, Lowenberg B, Touw IP. Distinct cytoplasmic regions of the human granulocyte colony-stimulating factor receptor involved in induction of proliferation and maturation. *Mol Cell Biol*. 1993;13(12):7774-7781.

24. Wouters BJ, Jorda MA, Keeshan K, et al. Distinct gene expression profiles of acute myeloid/T-lymphoid leukemia with silenced CEBPA and mutations in NOTCH1. *Blood*. 2007;110(10):3706-3714.

25. Hollink IH, van den Heuvel-Eibrink MM, Arentsen-Peters ST, et al. NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern. *Blood*. 2011;118(13):3645-3656.

26. Rockova V, Abbas S, Wouters BJ, et al. Risk-stratification of intermediate-risk acute myeloid leukemia: integrative analysis of a multitude of gene mutation and expression markers. *Blood*. 2011.

27. Pratcorona M, Abbas S, Sanders MA, et al. Acquired mutations in ASXL1 in acute myeloid leukemia: prevalence and prognostic value. *Haematologica*. 2012;97(3):388-392.

28. Ribeiro AF, Pratcorona M, Erpelinck-Verschueren C, et al. Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia. *Blood*. 2012;119(24):5824-5831.

29. Liu P, Keller JR, Ortiz M, et al. Bcl11a is essential for normal lymphoid development. *Nat Immunol*. 2003;4(6):525-532.

30. Wakabayashi Y, Inoue J, Takahashi Y, et al. Homozygous deletions and point mutations of the Rit1/Bcl11b gene in gamma-ray induced mouse thymic lymphomas. *Biochem Biophys Res Commun*. 2003;301(2):598-603.

31. Bernard OA, Busson-LeConiat M, Ballerini P, et al. A new recurrent and specific cryptic translocation, t(5;14)(q35;q32), is associated with expression of the Hox11L2 gene in T acute lymphoblastic leukemia. *Leukemia*. 2001;15(10):1495-1504.

32. Berger R, Dastugue N, Busson M, et al. t(5;14)/HOX11L2-positive T-cell acute lymphoblastic leukemia. A collaborative study of the Groupe Francais de Cytogenetique Hematologique (GFCH). *Leukemia*. 2003;17(9):1851-1857.

33. Przybylski GK, Dik WA, Wanzeck J, et al. Disruption of the BCL11B gene through inv(14)(q11.2q32.31) results in the expression of BCL11B-TRDC fusion transcripts and is associated with the absence of wild-type BCL11B transcripts in T-ALL. *Leukemia*. 2005;19(2):201-208.

34. De Keersmaecker K, Real PJ, Gatta GD, et al. The TLX1 oncogene drives aneuploidy in T cell transformation. *Nat Med*. 2010;16(11):1321-1327.

35. Gutierrez A, Kentsis A, Sanda T, et al. The BCL11B tumor suppressor is mutated across the major molecular subtypes of T-cell acute lymphoblastic leukemia. *Blood*. 2011;118(15):4169-4173.

36. Batanian JR, Dunphy CH, Gale G, Havlioglu N. Is t(6;14) a non-random translocation in childhood acute mixed lineage leukemia? *Cancer Genet Cytogenet*. 1996;90(1):29-32.

37. Bezrookove V, van Zelderen-Bhola SL, Brink A, et al. A novel t(6;14)(q25-q27;q32) in acute myelocytic leukemia involves the BCL11B gene. *Cancer Genet Cytogenet*. 2004;149(1):72-76.

38. Georgy M, Yonescu R, Griffin CA, Batista DA. Acute mixed lineage leukemia and a t(6;14)(q25;q32) in two adults. *Cancer Genet Cytogenet*. 2008;185(1):28-31.

39. Hayashi Y, Pui CH, Behm FG, et al. 14q32 translocations are associated with mixed-lineage expression in childhood acute leukemia. *Blood*. 1990;76(1):150-156.

40. Raimondi SC, Kalwinsky DK, Hayashi Y, Behm FG, Mirro J, Jr., Williams DL. Cytogenetics of childhood acute nonlymphocytic leukemia. *Cancer Genet Cytogenet*. 1989;40(1):13-27.

41. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell*. 2013;155(4):934-947.

42. Loven J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*. 2013;153(2):320-334.

43. MacLeod RA, Nagel S, Drexler HG. BCL11B rearrangements probably target T-cell neoplasia rather than acute myelocytic leukemia. *Cancer Genet Cytogenet*. 2004;153(1):88-89.

44. Cismasiu VB, Adamo K, Gecewicz J, Duque J, Lin Q, Avram D. BCL11B functionally associates with the NuRD complex in T lymphocytes to repress targeted promoter. *Oncogene*. 2005;24(45):6753-6764.

45. Wakabayashi Y, Watanabe H, Inoue J, et al. Bcl11b is required for differentiation and survival of alphabeta T lymphocytes. *Nat Immunol*. 2003;4(6):533-539.

46. Cherrier T, Suzanne S, Redel L, et al. p21(WAF1) gene promoter is epigenetically silenced by CTIP2 and SUV39H1. *Oncogene*. 2009;28(38):3380-3389.

47. Cismasiu VB, Duque J, Paskaleva E, et al. BCL11B enhances TCR/CD28-triggered NF-kappaB activation through up-regulation of Cot kinase gene expression in T-lymphocytes. *Biochem J*. 2009;417(2):457-466.

48. Cismasiu VB, Ghanta S, Duque J, et al. BCL11B participates in the activation of IL2 gene expression in CD4+ T lymphocytes. *Blood*. 2006;108(8):2695-2702.

49. Grabarczyk P, Nahse V, Delin M, et al. Increased expression of bcl11b leads to chemoresistance accompanied by G1 accumulation. *PLoS One*. 2010;5(9).

CHAPTER 11

Highly improved DNA copy number variation estimation from next generation sequencing data using reference data sets

Mathijs A. Sanders[1,*], Remco M. Hoogenboezem[1,*], Stefan Gröschel[1], Annelieke Zeilemaker[1], Wendy M.C. Geertsma-Kleinekoort[1], Ruud Delwel[1], Jelle J. Goeman[2,3,*] and Peter J.M. Valk[1,*]

[1] Erasmus University Medical Center, Department of Hematology, Rotterdam, The Netherlands
[2] Leiden University Medical Center, Department of Medical Statistics and Bioinformatics, Leiden, The Netherlands
[3] Radboud University Medical Center, Department for Health Evidence, Nijmegen, The Netherlands

* These authors contributed equally to this work

Submitted

ABSTRACT

A major cancer research field involves itself with the accurate detection of genetic lesions for the identification of novel cancer-related genes. Different next generation sequencing (NGS) platforms generate high-quality data enabling the estimation of copy number variations (CNVs). Various CNV estimation algorithms for NGS data have been developed, however, the ability to fully mitigate recurrent noise is still lacking. Therefore, a robust assessment of CNVs in high-variable regions, conferring recurrent statistical noise, is required. We developed a statistical framework utilizing information derived from diploid reference set samples and demonstrated a highly improved CNV estimation compared to contemporary methodologies. To exemplify the algorithm's strengths we applied our model on datasets derived from two distinct genome sequencing archetypes: (1) whole genome sequencing (WGS) data generated by Complete Genomics (CG) technology and targeted resequencing data generated on the Illumina HiSeq 2000, and (2) whole exome sequencing (WXS) data. The former embodies library strategies generating sequencing data from continuous regions, while the latter embodies a library strategy generating sequencing data from non-continuous and non-equidistant regions. To proof the effectiveness of our algorithm we compared our detected CNVs to those obtained with the Affymetrix DNA mapping arrays. We established that an increase in model resolution correlates with the number of detectable focal genetic lesions, largely corroborated by structural variants (SVs). Finally, our statistical framework produces good results for WGS and WXS data derived from different sequencing platforms, establishing its general applicability.

Supplemental material: http://hema13.erasmusmc.nl/mathijs_sanders/chapter11/

INTRODUCTION

Next generation sequencing (NGS) data enable the high-resolution characterization of cancer genomes.[1-4] In general, the purpose of sequencing genomic material of cancerous tissues is to uncover acquired mutations, short insertion-deletions (indels), and structural variants (SV). However, NGS reads can also be used for the estimation of copy number variations (CNV). In fact, high-density sequence reads enable the detection of genetic lesions serving as guides towards the identification of novel cancer-related genes.[5-9]

Estimating CNVs is particularly difficult within high-variable regions. These regions mainly comprise repeat elements (e.g. LINEs, SINEs, and LTRs) which are notorious genomic elements conferring improper or ambiguous aligned reads. Additionally, systematic bias inherent to the individual sequencing technologies result in specific and incorrect fragment count patterns, presenting yet another unanswered challenge. These biases result in a consistent number of regions with under, but mostly, over-estimated copy number estimations observable in all sequenced samples. Additionally, most CNV estimation algorithms provide a spatial resolution too large for robust detection of genetic lesions affecting small genetic elements, e.g. exons. For example, the Complete Genomics (CG) pipeline provides a spatial resolution of 100 kb for sequenced tumor cases, which is sufficient for detecting large genetics lesions, however, is too coarse for revealing genetic lesions smaller than the spatial resolution.

To mitigate these biases and ameliorate the quality of CNV detection, we developed a statistical framework for CNV estimation from NGS data by integrating information derived from reference set samples. *De facto*, the high-variable regions are located on the same physical position in different individuals, although with different fragment count magnitudes (Supplementary Figure 1), enabling the extraction of these intrinsic characteristics from the fragment reference sample count profiles. These characteristics are used to filter and normalize the noisy fragment count profiles from samples of interest, ultimately, producing filtered CNV profiles which could serve as input for further analyses.

We demonstrate the improved CNV estimation procedure by applying it to three distinct datasets derived from two genome sequencing archetypes. The first archetype embodies technologies sequencing continuous regions, implying that a particular region of the genome is completely sequenced without non-sequenced interfering regions. For this archetype context we applied the model on whole genome sequencing (WGS) data generated from 3 acute myeloid leukemia (AML) cases and matched controls by CG DNA nanoball sequencing technology. Subsequently, we utilize our model on targeted resequencing data from the chronic myeloid leukemia (CML) cell line K562, generated by the Illumina HiSeq 2000 platform. For the second sequencing archetype context, we applied the model on whole exome sequencing (WXS) data derived from 6 unrelated *de novo* AMLs with matched controls and matched relapse samples. To assess the model sensitivity we compared the determined CNVs to those obtained with Affymetrix DNA mapping arrays. We established that an increased model resolution leads to a

gain in detectable CNVs and validated the veracity of these CNVs by corroborating them with the independent *in silico* methodology of structural variants (SVs). In conclusion, we demonstrated that the algorithm enables the prediction of CNVs using diverse NGS platforms.

MATERIAL AND METHODS

Package

We compiled all described methods in the package *CNVsvd*, freely available from http://hema13.erasmusmc.nl/CNVsvd/. Our package is a C++ (gcc 4.7.2) and R (R 2.15.2) package that extracts coverage information from appropriate source materials – coverage information files for CG sequencing data and BAM files for other sequencing platforms – according to user preferences and generates CNV profiles ready for importation into SNPExpress[10], included in the package, or CNV segmentation algorithms[11,12] for further processing. The package is tested on different operating system platforms (Linux: Ubuntu 10.04, Windows: Windows XP and 7).

Sample processing

Complete genomics

Files containing coverage information for 45 healthy reference set samples were downloaded from the CG website (Complete Genomics Diversity set repository[13], NCBI 36, Pipeline 1.10.0). Our 3 CG AML samples and their matched controls were processed using NCBI 36 and Pipeline version 1.10.0. Consequently, the reference set and AML samples are directly comparable. Affymetrix 500K DNA mapping arrays for the same 3 AMLs were processed as reported previously.[10]

Targeted resequencing

Targeted resequencing data of chromosomal regions 3q21 and 3q26 was generated by targeted custom capture beads (Nimblegen v2) for 8 AML samples (cell line K562, and 7 normal karyotype AMLs [NK-AMLs]) following the manufacturer's protocols. Captured samples were sequenced on the Illumina HiSeq 2000 and aligned against human genome 19 (hg19) using the Burrows-Wheeler Aligner 0.5.9.[14] Coverage information per window was extracted by using the SAMtools API 0.1.13[15], included in the software package. Structural variants were determined by BreakDancer 1.1.[16]

Whole exome sequencing

WXS sequencing data was generated from exome bead captured material (SeqCap EZ Human Exome Library v3.0[17]) for 6 *de novo* AML cases (Supplementary Table 1), 6 matched *in vitro* expanded T-cell controls, 4 matched relapse AML cases out of the 6, and 24 unrelated *in vitro* expanded T-cell controls functioning as an addition to the reference dataset. Captured samples

were sequenced on the Illumina HiSeq 2000 and aligned against hg19 using BWA. We developed an algorithm that utilizes a provided BED file containing the exon interval structure and extracts the fragment counts accordingly. The fragments are extracted by using the SAMtools API.[15] Affymetrix 250K Styl DNA mapping arrays were processed as previously reported.[10]

In vitro T-cell expansion

T-cells were isolated from the diagnostic material of each AML case and expanded with anti-CD3 and anti-CD28 monoclonal antibodies in the presence of IL-2. After 2 weeks T-cells were enriched with MACS separation columns. All diagnostic material derived T-cell populations were shown to be >98% pure and served as normal controls.

Window estimations

Continuous sequenced regions

Fragment count statistics were determined from coverage information within consecutive non-overlapping windows positioned along the genome. The window size, i.e. resolution, is determined based on the application of desire, e.g., detecting small CNVs with higher resolution. Too large windows could result in missing small focal aberrations, while too small windows might yield substantial fragment count variability. Coverage information from all nucleotide positions within a given window contribute equally to the count statistic. We removed all windows with a consistent low or zero coverage in all reference set cases to improve numerical stability. Population variation is a strong confounder for estimating CNVs in particular genomic loci due to population divergence. Options in the algorithm are included to omit regions devoid of reads in a fraction of the reference set to avoid population variation confounding. Due to low numbers of male references samples, we excluded chromosome Y and multiplied the coverage information of all windows within chromosome X by two for male samples.

Non-continuous sequenced regions

Given the non-continuous nature of WXS we extracted the fragment counts according to the exon structure. Fragment counts are extracted from BAM files by utilizing a BED file containing the captured region intervals. The algorithm requires, as an input, the desired size of the window ω (Figure 1). Captured fragments originating from interval edges normally extend beyond the interval, therefore we included a parameter σ extending the interval edges enabling the extraction of full fragments. Fragments are only considered when uniquely aligned and count estimates are stratified on map quality. Unpaired reads are retained only when aligned uniquely with a sufficient mapping quality. Consecutive non-overlapping windows are used when an exon is sufficiently large to encompass multiple windows. If a portion of the exon remains uncovered by a window, i.e. due to the window size, a new window is introduced if this portion is larger than half of the window size; otherwise it is added to the adjacent window.

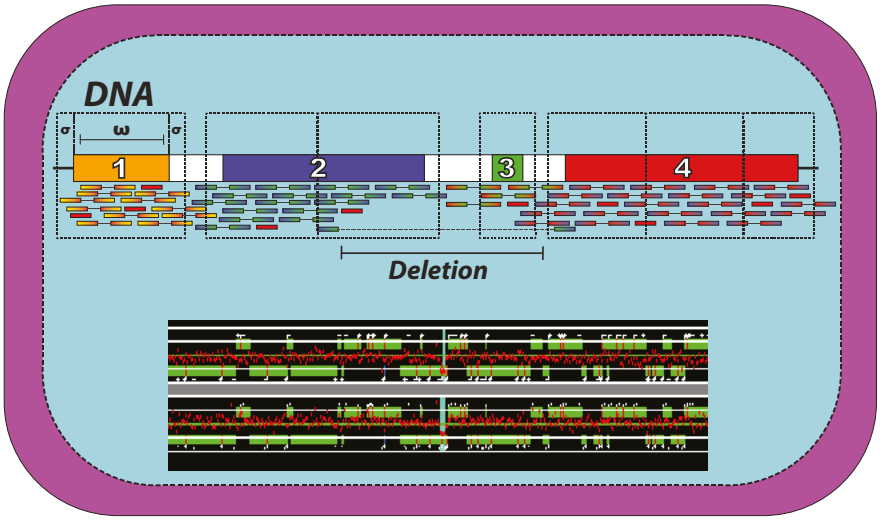


Figure 1. Generating CNV profiles from WXS data. Specific regions of the genome are captured, e.g. exome. The algorithm determines based on the window size ω, the extension parameter σ and exome structure how to optimally place the windows. For each window the fragment counts are determined. Within the depicted locus a region, encompassing partially exon 2 and completely exon 3, is deleted resulting in a lower number of fragment counts. The deletion also results in a SV depicted by a paired-end read with discordant distance. The fragment counts are processed by the algorithm and compared to the reference set. The resulting CNV profile is illustrated in SNPExpress[10] and the same deletion is visualized for two patients (baby blue).

Normalization

Supervised quantile normalization

All resulting count profiles derived from continuous regions were normalized by supervised quantile normalization.[18] Reference set samples were used to construct a reference distribution and all samples, including AML samples, were normalized against this distribution.

Adapted FPKM

We developed a normalization technique based on the fragments per kilobase of exon per million fragments mapped (FPKM) statistic. Given the non-continuous exome structure and differing window sizes the FPKM statistic would better suit the nature of WXS. The adapted FPKM is calculated for every window by:

$$FPKM_i = \frac{F_i}{L_i * N}$$

Where $FPKM_i$ is the FPKM statistic, F_i is the fragment count, and L_i the length for window *i*. Finally, the statistic is normalized by the total number of counted fragments *N* passing the quality criteria.

Scaling

Subsequently, the fragment statistics are scaled for further processing. First, the mean fragment count for each window was computed from the reference set samples. For all samples, including the reference set samples, the fragment counts were divided by the calculated mean per window. If the genomic locus of the current window is not afflicted by a genetic lesion, the resultant statistic would equate to 1.

Singular value decomposition of the reference samples

We assume that for reference set samples, e.g. healthy individuals, on average the copy number equals 2. Commonly healthy individuals harbor many different natural CNVs[19,20], nevertheless, we assume that these are not recurrent in the majority of the reference set samples. Given these assumptions we hypothesized that recurrent variation stems from high-variable regions and systematic biases inherent to the sequencing technology. The high-variable regions share the same genomic loci for all sequenced cases, irrespective of disease background (Supplemental Figure 1 and Figure 2A). Interestingly, this variation is described by the first few principal components derived from the reference set fragment count profiles, called systematic noise components. We found that typically the first two components contain most of the systematic variance (data not shown), however, an appropriate number of components can be determined from the singular values.[21] Utilizing singular value decomposition (SVD) for noise eliminating is an established methodology.[22] Systematic noise removal by principal components for AML cases harboring monosomies, i.e. loss of chromosomes, or a complex karyotype, i.e. a mixture of chromosomal losses and gains, presents a difficult context. Numerical changes of chromosomes is prevalent among different cancer subtypes and strongly associated with prognosis and treatment outcome, e.g. AML.[23] Utilizing principal components, without appropriate preprocessing, results in the obscuration of chromosomal losses or gains. This conundrum is solved by calculating a mean statistic from the normalized count profile for every chromosome of a sample. This mean statistic is subtracted from the normalized profile per chromosome resulting in a count profile centered around 0. These statistics are preserved to be re-added after noise removal. This procedure is viable as it does not interfere with SVD or noise removal. First, we calculate:

$$\vec{\boldsymbol{y}}_{\boldsymbol{i,j,centered}} = \vec{\boldsymbol{y}}_{\boldsymbol{i,j}} - \boldsymbol{\mu}_{\boldsymbol{i,j}} \; (\mathbf{1.1})$$

Where $\vec{y}_{i,j,centered}$ is the normalized count profile centered around 0 for sample *i* and chromosome *j*, $\vec{y}_{i,j}$ is the normalized count profile and $\mu_{i,j}$ the mean statistic of the count profile for sample *i* and chromosome *j*. For every chromosome the normalized data of the reference set samples are structured into a data matrix X (n samples x m covariates, n < m). Using SVD we have the following:

$$\boldsymbol{X} = \boldsymbol{U\Sigma V^T} \; (\mathbf{1.2})$$

Here U (n x n) and V (n x m) are matrices with orthogonal basis vectors. The matrix Σ (n x n) contains the singular values. The matrix V contains the systematic noise components, describing variance observed in the reference set while the singular values amount to the variance strength.

For SVD it is important not to include the centered count profiles of cancer samples into X. Including cancer samples runs the risk of removing components describing recurrent genetic aberrations seen in the disease of interest, especially if the reference data set is small relative to the cancer data set. To determine the systematic noise components from these large matrices would be computational intensive. Therefore, we combined eigenvalue and singular value decomposition:

$$\boldsymbol{XX^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = UDU^T} \ (\mathbf{1.3})$$

The data matrix given in (equation 1.3), is only of size n x n (i.e. 45 x 45 for the CG data, 7 x 7 for the targeted resequencing, and 30 x 30 for the WXS data), and enables the fast computation of the matrices D and U. We subsequently determined the noise component matrix V:

$$\boldsymbol{V^T = \sqrt{D}^{-1} U^{-1} X} \ (\mathbf{1.4})$$

Principal component regression

SVD determines *n* high-dimensional noise components per chromosome. We used the *g* strongest noise components in a principal component regression procedure estimating their contribution to the count profiles of the AML and matched control samples:

$$\boldsymbol{\hat{y}_{i,j,noise} = \beta_1 \vec{v}_{j,1} + \cdots + \beta_g \vec{v}_{j,g}} \ (\mathbf{2.1})$$

The systematic noise present in the fragment count profile of sample *i* and chromosome *j* is modeled by $\hat{y}_{i.j.noise}$, determined by fitting the systematic noise components to the centered fragment count profile (Figure 2A). Ultimately, we remove the estimated noise contribution from the normalized data and re-add the mean statistic to preserve the possibility to detect whole chromosome aberrations:

$$\boldsymbol{\vec{y}_{i,j,filtered} = \vec{y}_{i,j,centered} - \hat{y}_{i,j,noise} + \mu_{i,j}} \ (\mathbf{2.2})$$

The resultant filtered fragment count profiles contain residual sample specific noise and copy number variations.

Copy number estimation

Principal component regression is effective for noise removal in high variable regions, however, it cannot reduce sample specific noise. Therefore we performed spatial filtering (median smoothing) for applications involving continuous regions, trading some spatial resolution for less noisy coverage estimation. Finally, the filtered residuals are linearly transformed to represent the correct copy number variation estimations (Figure 2B). Due to the non-equidistant spacing of the exome, median smoothing has not been applied on WXS data. Instead, for WXS we considered a genetic aberration detected if three adjacent windows with similar CNV estimates were observed.

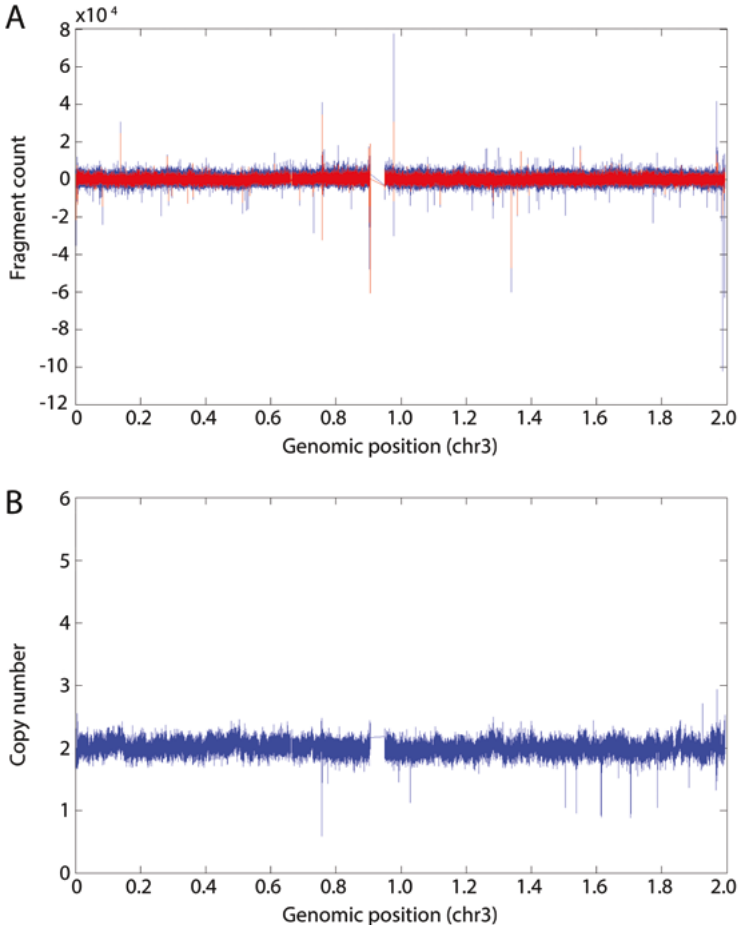


Figure 2. Fitting noise components to the count profile. (Blue) Count profile for chromosome 3 was generated with a 0.5 kb resolution from CG data (Red). Noise profile was determined by fitting the noise components to the count profile. (A) First noise component fitted to the count profile of an AML case. Of note, some blue peaks are almost completely covered by the noise profile. (B) Estimated CNV profile after normalization and noise removal of the same AML case.

DATA DESCRIPTION AND VALIDATION

Complete genomics data set: continuous sequenced regions

(I) Comparing to CG CNVs

We obtained 3 unrelated AMLs with matched controls sequenced with CG's DNA nanoball sequencing technology. Initially, we compared the CNVs detected by our algorithm with different window sizes, 5 and 10 kb, to those detected with CG's pipeline which uses 100 kb windows for tumor samples.

(II) Comparing to DNA mapping arrays

To demonstrate the sensitivity of our algorithm we characterized all 3 AMLs on the Affymetrix 500K DNA mapping array. Subsequently, we compared CNVs detected with our algorithm, yielded with a 5 kb window size, to those detected with the array.

(III) Increased resolution within the algorithm

To infer if an increase in the resolution of the algorithm yielded a gain in CNV detection sensitivity, we compared CNVs detected with a window size of 5 kb those obtained with a window size of 0.5 kb.

(IV) Corroboration with structural variants

The veracity of detected CNVs, detected with a 0.5 kb window size, is assessed by corroboration with SVs. This *in silico* methodology independently detects genetic lesions and provides further evidence. SVs are mainly used for the detection of deletions, large inter- and intrachromosomal events (e.g. translocation, inversions), and tandem duplications, however, is not used for detecting amplifications or complex genetic aberrations. The nature of the SVs prevents the detection of all amplification events, resulting in our primary focus on deletion CNVs, i.e. copy number of 0 or 1, for corroboration with SVs.

(V) Somatic CNVs

Finally, assessing the somatic status of CNVs (sCNV) is essential for cancer research. The matched control is used to determine the number of somatic CNVs detected with a 0.5 kb window size.

Targeted resequencing data set: continuous sequenced regions

(I) CNV estimation

We determined CNVs in the CML cell line K562 from targeted resequencing of the 3q21 and 3q26 chromosomal loci. In total 7 NK-AMLs were utilized as reference set samples. CNV detection was performed by setting the window size to 0.5 kb. SVs were detected and used as corroborating evidence for the determined CNVs.

Whole exome sequencing data set: non-continuous sequenced regions

(I) CNV estimation

CNVs were determined from the WXS dataset with a 0.5 kb window size, an extension parameter of 0.1 kb, and utilizing a BED file describing the exon structure. The data set comprises 6 diagnostic *de novo* AML with 4 matched AML relapse samples and for all samples a matched T-cell control, implying that 2 *de novo* AML cases did not have matched relapse samples. The reference set included 24 unrelated T-cell control samples derived from other AML patients and the 6 matched T-cell control samples.

(II) Comparing to DNA mapping arrays

All 6 *de novo* AML samples were characterized on Affymetrix 250K *Sty*I DNA mapping arrays. We compared the CNVs detected by our algorithm to those obtained with the DNA mapping arrays.

(III) Conundrum and validation

Validating CNVs in WXS data is challenging. Unless the breakpoint occurs directly in an exon or its proximity, there are no structural variants validating the determined CNVs. To resolve this issue detected CNVs in the AML samples were deemed true if it was also detected in the relapse samples, but not in the T-cell controls, i.e., a sCNV, or in both the relapse and control samples, i.e., a germline CNV.

RESULTS

We investigated the robustness of the newly developed algorithm within two genome sequencing archetypes. Each sequence archetype presents its own set of conundrums addressed by the novel statistical framework irrespective of the sequencing platform. Using three data sets representing the two archetypes we highlight examples and aspects exemplifying the strength and adaptability of the algorithm.

CNV estimations in continuous sequenced regions

Complete Genomics DNA nanoball sequencing technology

We demonstrate the algorithm sensitivity and exactness of the detected CNVs by using different comparisons. The results are highlighted by a single index AML patient, AML #1. The results for the remaining AML cases, AML #2 and #3, are listed in the tables or in the supplementary material whenever noted.

Comparison to the CG pipeline

We appraised the efficiency of noise estimation and removal by comparing the CNVs determined by the CG pipeline to those obtained with our algorithm with 5 and 10 kb window sizes. Genome-wide CNV profiling by the CG pipeline revealed a number of regions with medium to very high copy numbers (Figure 3A) observed consistently in the AML as well as reference samples (data not shown), comprising high-variable or pericentromeric regions, impairing robust CNV estimation. Our model mitigates this intrinsic bias, thereby decreasing the contribution of systematic noise to the AML CNV profiles (Figure 3B an 3C). Interestingly, the model revealed additional focal genetic lesions, smaller than the spatial resolution of the CG pipeline (orange box Figure 3, detailed in Supplementary Figure 2). Irrespective of the window size, 5 or 10kb, similar CNV profiles were estimated with only some additional genetic lesions detected using the smaller window size, indicating that an increased resolution correlates with CNV detection efficiency (Figure 3B).

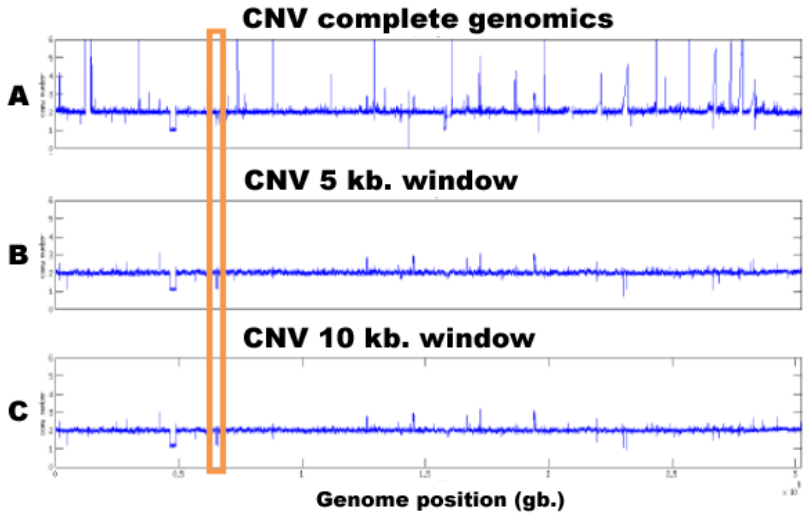


Figure 3. Copy Number estimations for the complete genome of one AML case. (A) CG CNV profile estimated with 100 kb windows. (B) Copy number estimations from consecutive 5 kb windows. (C) Copy number estimations from consecutive 10 kb windows. (Orange box) Some deletions detected by our algorithm, but not by CG's pipeline. Regions shown in more detail in Supplementary Figure 2.

CNV estimation is essential for delineating genetic events

Enhanced CNV estimation provides a beneficial feature for disease delineation. Not all genetic lesions are detected by SVs and enhanced CNV estimation could provide further evidence for genetic lesion acquisition. For example, the AML index patient revealed copy number gains encompassing the gene *BCR* on chromosome 9 and *ABL1* on chromosome 22 (Supplementary Figures 3 and 4), formed by an unbalanced translocation, resulting in the *BCR-ABL1* fusion

transcript, a tell-tale marker of CML, missed by routine cytogenetics. We confirmed the presence of the *BCR-ABL1* fusion transcript in AML#1 by RT-PCR (Supplemental Figure 5). The translocation was detect by SVs, however, the gain of genetic material on each side of the breakpoint remained undetected.

Detecting additional CNVs with respect to DNA mapping arrays

DNA mapping arrays are the traditional methodology for estimating CNVs. Probe sets designed to measure single nucleotide polymorphism (SNP) genotypes are also utilized for CNV estimation. We compared the CNVs detected by our algorithm to those obtained with the DNA mapping array and determined the overlap and differences (Table 1A), revealing that our algorithm detects substantially more CNVs (e.g., Figure 4). CNV analysis on the index AML case demonstrated that the algorithm detected all genetic lesions observed with DNA mapping arrays and more (Table 1A and Supplementary Figure 6). Interestingly, not all CNVs detected by DNA mapping arrays for the remaining two AML cases were detected by our algorithm. Diligent visual inspection of the missed CNVs revealed that these are presumably false positives, given the local pattern of CNV profile variance (e.g. intermittent gains and losses at the same genomic locus), or are located at regions difficult for reliable CNV estimation (e.g. pericentromeric or high-variable regions). The latter case demonstrates that the principle component procedure attenuates the effect of these regions on the CNV profile estimation.

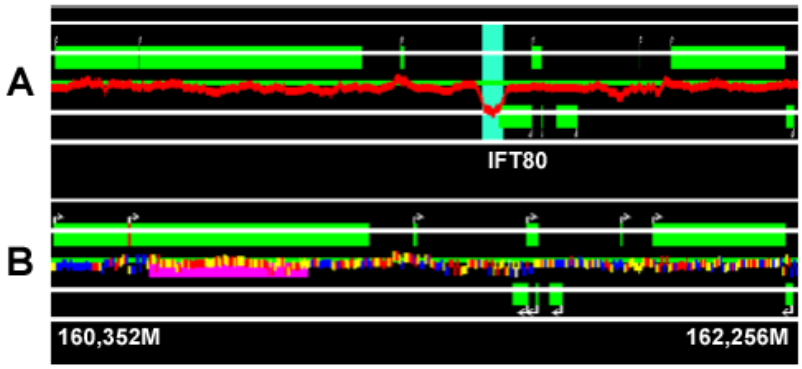


Figure 4. CNV on chromosome 3. (A) Deletion of a region (baby blue) on chromosome 3 for AML sample #1 detected by CNV estimates from our algorithm utilizing CG data. (B) This deletion is not detected when AML sample #1 was analyzed using the Affymetrix 500K DNA mapping array.

Increased resolution correlates with enhanced CNV detection

An increased resolution of the CNV estimation algorithm enables the detection of smaller genetic lesions, which could play a pivotal role in the pathogenesis of leukemia.[24] While large genetic lesions frequently occur in cancerous tissues, small and specific genetic lesions enable the

determination of novel cancer-related genes, especially in regions harboring CNVs of different sizes. Focal aberrations substantially smaller than the window size will not confer a large difference in the total fragment count and are therefore likely to be missed.

Increasing the algorithm's resolution from 5 kb to 0.5 kb enhanced the CNV detection rate. Most additionally detected CNVs were of the size 1.5-6 kb, implying that a 5 kb resolution is insufficient for detecting lesions. In the AML index patient, all but a single CNV identified with a 5 kb resolution were also identified with a 0.5 kb resolution (98% for AML #1, Table 1B). Similar patterns were observed in the other two AML cases (68% and 80% for AML #2 and AML #3 respectively). CNVs remaining undetected with the higher resolution, 0.5 kb, but detected with the lower resolution, 5 kb, are mostly associated with regions encompassing multiple high-variable regions increasing the fragment count variance substantially. The increased granularity, due to a higher resolution, enables the delineation of local systematic noise signatures into separate components.

Structural variants corroborate the detected CNVs

WGS enables the determination of structural variants which are utilized for determining indels, translocations, and inversions.[25] Paired-end reads derived from homologous regions, e.g. repeat regions and pseudogenes, are the leading cause for detecting false positive SVs, if aligned to ambiguous positions. CNV estimations could serve as independent corroborating evidence for deletion events, further increasing the detection likelihood for true genetic lesions.

Increasing the model resolution led to a gain in CNVs detected, however, the question remains if they represent true genetic lesions. We determined that most CNVs detected by a 5 kb resolution are corroborated by SVs increasing the likelihood that these CNVs are true genetic lesions. Subsequently, we determined that the majority of detected CNVs with a 0.5 kb resolution case are corroborated by SVs (Table 1C). On average, 60% of additionally detected CNVs are corroborated by SVs. Careful visual inspection confirmed that most non-corroborated CNVs are likely true genetic lesions. Strikingly, these CNVs are supported by multiple consecutive windows with consistent similar copy number estimates. In some instances, the same CNV is detectable in the matched control sample hinting at a germline CNV not detected by SVs. Likewise, analyses from the other 2 AMLs demonstrated a similar pattern, a substantial proportion of CNVs are corroborated by SVs (Table 1C). Similar to the index patient the non-corroborated CNVs are strongly supported by CNV estimates from multiple consecutive windows (Supplementary Figure 7).

Somatic CNVs

CNV profiles for matched controls where generated, i.e from *in vitro* expanded T-cells, by applying the same statistical framework, enabling the determination of sCNVs. Surprisingly, most of the CNVs detected in the three AML cases appeared germline CNVs as they were detected in the matched control. In the index AML patient only 10 CNVs were deemed somatic (Table 1D), which comprised small (1.5 – 6 kb) and large CNVs. For the other two AML cases the number of sCNVs was even lower.

Table 1. Detection of CNVs in 3 AML cases characterized by NGS and DNA mapping arrays.

	A. SNP array - 5kb. windows			B. 5kb. - 0.5 kb. windows			C. Detected by SV (deletions)		D. 0.5 kb
	Array	5 kb.	Overlap	5 kb.	0.5 kb.	Overlap	0.5 kb.	Overlap	Somatic
AML #1	**9**	**42**	**9(100%)**	**42**	**121**	**41(98%)**	**80**	**54(68%)**	**10(8%)**
AML #2	**5**	**31**	**3(60%)**	**31**	**112**	**21(68%)**	**81**	**48(59%)**	**2(2%)**
AML #3	**3**	**44**	**1(33%)**	**44**	**110**	**35(80%)**	**66**	**44(67%)**	**8(7%)**

(A) Novel against traditional. The NGS algorithm utilizing CG data with 5kb resolution in comparison to the Affymetrix 500K DNA mapping array. (B) Resolution. CNVs detected by increasing the resolution from 5 kb to 0.5 kb. (C) Intersection of CNVs detected with a resolution of 0.5kb and the detected SVs (D) sCNVs detected in the tumor sample.

Targeted resequencing of predefined genomic loci

We investigated whether the algorithm is applicable to sequencing data derived from different sequencing platforms, e.g. Illumina HiSeq 2000. The CNV profile for the K562 cell line was determined by utilizing a reference set comprising 7 NK-AMLs. In total, 7 CNVs, i.e. 6 tandem duplications and 1 focal deletion, were detected and mainly conferred by tandem duplications as evidenced by SVs. For example, an amplification involving the gene *PPM1L* was detected (Supplementary Figure 8 and 9) and revealed by SV analysis to be a consequence of a tandem duplication.

Additionally, our algorithm detected, in conjunction with a corroborating SV, a specific tandem duplication within the gene *MECOM* (*MDS1-EVI1* complex locus), encoding for MDS1 and EVI1 (Supplementary Figure 10). The gene *EVI1* encodes a proto-oncogene which, upon overexpression, confers a dismal prognosis in AML.[26,27] This yet unreported genetic lesion may explain *EVI1* overexpression, previously reported for K562.[28] The identified tandem duplication was validated by Sanger sequencing (Supplementary Figure 11).

Conclusion

Ultimately, these data demonstrate that our analysis tool is applicable to sequencing data derived from different sequencing platforms. Both sequence methodologies, i.e. CG DNA nanoball sequencing and Illumina HiSeq 2000, demonstrate systematic noise in the fragment count profiles, which is substantially mitigated by our algorithm. We have established its general applicability, increased sensitivity, and overall increased specificity on data derived from continuous sequenced regions.

CNVs estimations in non-continuous sequenced regions

Whole exome sequencing

Sequencing of non-continuous and non-equidistant regions is difficult for robust CNV estimation due to the mainly unrelatedness of adjacent windows. Although WXS data is less affected by high-

variable regions, located mainly in intergenic and intronic regions, it is still perturbed by other inherent properties of the capture design, e.g. the capture of pseudogenes.

Detecting CNV aberrations

We next investigated the robustness of the newly developed algorithm for CNV detection in non-continuous sequencing data. First, we demonstrated that our algorithm is able to detect whole chromosomal aberrations, therefore accommodating the detection of aneuploidy. A gain of chromosome 8 was detected in the leukemic blasts of patient 2215 corroborated by a DNA mapping array and standard cytogenetics (Supplementary Figure 12A, Supplementary Table 1). Of note, cytogenetics denoted that this gain was present in 60% of the cells, implying that the algorithm detects subclones. An acquired gain of chromosome 11 was detected in patient 2226, corroborated by a DNA mapping array and cytogenetics (Supplementary Figure 12B). Finally, multiple CNVs of differing sizes have been detected for all 6 AML patients (range: 7-21, Figure 5).

Comparison to DNA mapping arrays

All 6 *de novo* AML cases were characterized by DNA mapping arrays. Careful inspection revealed that the algorithm was able to corroborate all CNVs detected by DNA mapping arrays, except for a single CNV encompassing the pseudogenes *IGLL3P* and *LRP5L*. Visual inspection of DNA mapping array data revealed that this region remains difficult for robust CNV estimation (data not shown), as it comprises multiple pseudogenes and could be equated to high-variable regions. The effect of this region on the fragment count profile was mitigated by the noise removal methodology within the algorithm.

AML patients have an exceptionally low number of CNVs, with a mean of 2.38 CNVs per case as determined by DNA mapping arrays.[29] Strikingly, compared to the DNA mapping arrays, the algorithm detected additional CNVs in all 6 AML cases. Interestingly, most additionally detected CNVs are corroborated by the matched control or relapse samples. We observed a significant increase in CNVs detected by our algorithm when compared to DNA mapping arrays (p=0.0345, Figure 5).

The detection of the clinically relevant *MLL*-PTD

Initially, a gain was detected in the gene *MLL*, located on chromosome 11, in patient 2226. This patient harbors a gain of chromosome 11, however, a specific region within *MLL* is additionally amplified and not observed in the matched control or relapse samples (Supplementary Figure 13). Given current knowledge, this implies that the patient acquired a *MLL* partial tandem duplication (*MLL*-PTD), which remained undetected by DNA mapping arrays. The aberration *MLL*-PTD comprises a duplication of consecutive exons within the gene *MLL* and is postulated to result in a recessive gain-of-function[30] and to associate with shortened remission time.[31] The partial tandem duplication is hard to detect by standard PCR methodology as there is no consensus on which

consecutive exons are duplicated in tandem, necessitating multiplexed PCR, which occasionally produces false positives. The algorithm enables the detection of this aberration, which was corroborated by a SV, demonstrating that the *MLL*-PTD was acquired (Supplementary Figure 14).

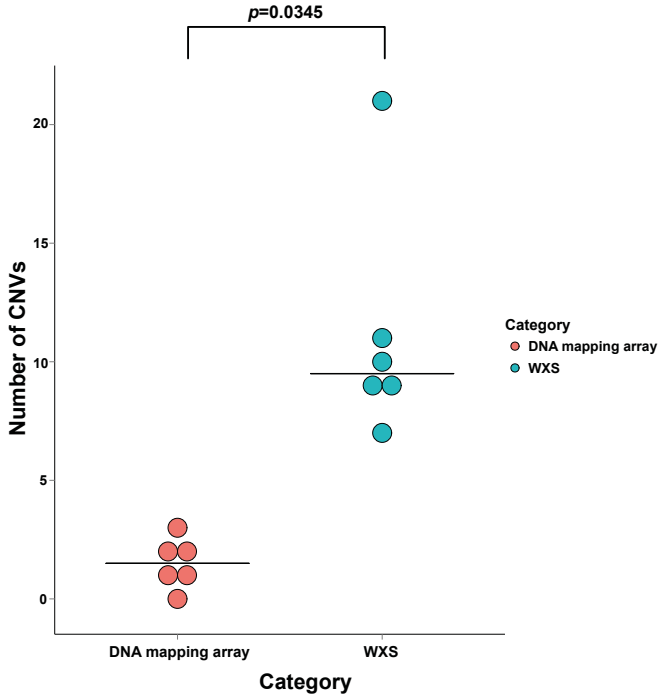


Figure 5. Number of CNVs detected with DNA mapping arrays and WXS. (Pink) Number of CNVs detected in 6 *de novo* AML cases by DNA mapping arrays (Blue) Number of CNVs detected in the same 6 *de novo* AML patients using WXS data. *P*-value was determined with the Wilcoxon rank-sum test.

DISCUSSION

We developed a statistical model enabling improved and reliable CNV estimation from NGS methodologies by removing systematic variation introduced by high-variable regions. The optimal choice of systematic noise components depends on the pervasiveness and correlation structure of the systematic noise. A scree plot, depicting the persistence of the noise components through eigenvalues, facilitates the choice of the optimal number of noise components.[21] Choosing a too low number of components results in the retention of systematic noise, while there is no heuristic for choosing too many components. Systematic noise is mainly contained within the first few noise components, while the remaining components mainly comprise variance conferred by natural CNVs. These remaining noise component will not optimally fit to count profiles if cases of interest do not harbor these natural CNVs, thereby conferring the regression coefficient to

become approximately zero and omitting the noise component from contributing to the noise estimation. Interestingly, for WXS sequencing applications, only one principal component seems sufficient, contrasting to WGS data for which at least two components are necessary. We advance the hypothesis that this is due to the nature of the exome, containing little to no repeat sequences. Systematic noise in WXS data mainly arises from systematic technical noise and the accidental capture of pseudogenes. Using normalization in combination with noise reduction we were able to generate high-resolution CNV estimations from NGS coverage information. Essential for our methodology is the presence of a reference data set of diploid controls, used for normalization and determination of the systematic noise components.

In comparison, the WGS based CNV estimation algorithm *CNVnator*[5] uses fragment count information from tumor and normal pairs, however, is unable to reduce systematic variance as it does not utilize variance information derived from a reference set. *Varscan 2*[7], a WXS based CNV estimation algorithm, uses fragment count information from tumor and normal pairs to estimate CNVs, however, it does not provide methods for dealing with high-variable regions. *CoNVEX*[32], specifically designed for WXS data, uses discrete wavelet transformation (DWT) for noise reduction and estimates the CNVs using a Hidden Markov Model (HMM). It uses average read depth ratios from tumor and matched normal as input. Although the DWT will reduce noise in some instances, it will not be able to: (I) calculate the absolute CNV, only denoting if regions are gained or lost, (II) capture the normal population fragment count variance. The latter results in the retention of high-variable regions, especially when not present in the matched normal control. *CoNIFER*[33], specific for WXS based CNV estimation, closest resembles our algorithm. Likewise, it employs SVD to normalize count profiles of samples of interest. Fundamental differences lay in the normalization, derivation of the SVD and the use of its singular value components. *CoNIFER* derives the SVD from all sequenced samples simultaneously, i.e., from a mix of cases and controls. This methodology is valid only under the assumption that there are no recurrent genetic aberrations in the sequenced samples of interest, which could contribute to a strong systematic noise component. Sequencing experiments involving cancerous tissues with similar genetics lesions provides an example which invalidates this assumption. *CoNIFER* generally classifies these aberrations as noise, subsequently removes them, and is therefore unable to detect whole chromosome aberrations in aneuploidic samples. Known examples of whole chromosomal aberrations involve chromosome 5 and 7 in AML, which have been associated with poor prognosis.[34,35] Missing those aberrations would omit vital information for correct classification or prognostication of these AML cases. Similar to our algorithm, *CoNIFER* normalizes the fragment count statistics to make them comparable. However, unlike our algorithm, *CoNIFER*'s Z-RPKM loses the ability to detect absolute copy number without a large set of additional samples. Knowing the absolute copy number could be essential for prognostication and delineating the pathogenesis of the disease of interest.[36,37]

The algorithm enables CNV estimation on data generated from different library strategies. We demonstrated, based on a CG data set, that the algorithm detects more validated CNVs than

provided by CG's native algorithm and at a higher resolution. Additionally, the comparison of high-density DNA mapping arrays to our algorithm demonstrated the superiority of the latter to detect novel CNVs. We have increased the detection resolution as far as 0.5 kb and compared it to a statistically more robust, but less detailed resolution of 5 kb. The algorithm detected additional CNVs, mostly corroborated by SVs, with sizes as small as 1.5 kb. Subsequently, we detected many tandem duplications in the targeted resequencing data of the cell line K562 which were corroborated by SVs. We demonstrated a tandem duplication within the *MECOM* locus, which could result in the overexpression of the dismal marker *EVI1*.

In addition we demonstrated that our algorithm is applicable to sequencing data from non-continuous regions. We demonstrated that the algorithm, using WXS data, is able to detect all CNVs detected with DNA mapping arrays except one. Interestingly, the algorithm detected more CNVs compared to DNA mapping arrays, which is striking as consensus dictates that AML has a very low frequency of genetic aberrations. Additionally, we were able to detect aberrations leading to partial tandem duplications in the gene *MLL*, which were not detected by the DNA mapping arrays.

In conclusion, we demonstrated that our novel statistical framework reliably detects more focal CNVs in different types of NGS data which is of great importance for detecting genetic abnormalities underlying the disease of interest.

Acknowledgments

This study was performed within the framework of CTMM, the Center for Translational Molecular Medicine Leukemia BioCHIP project (grant 03O-102).

Authorship

MAS: Developed statistical model, analyzed data and wrote manuscript; RH: statistical model, analyzed data; SG, AZ, WCMGK: Isolated, sequenced and analysed the samples; RD: Intellectual contribution and provided samples; JJG: Intellectual contribution to statistical model and wrote manuscript; PJMV: Designed research, analyzed data and wrote manuscript. The authors of this manuscript have nothing to disclose.

REFERENCES

1. Chapman MA, Lawrence MS, Keats JJ, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*. 2011;471(7339):467-472.
2. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014.
3. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214-218.
4. Welch JS, Ley TJ, Link DC, et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*. 2012;150(2):264-278.
5. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011;21(6):974-984.
6. Alkan C, Kidd JM, Marques-Bonet T, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet*. 2009;41(10):1061-1067.
7. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-576.
8. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet*. 2013;45(10):1134-1140.
9. Groschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369-381.
10. Sanders MA, Verhaak RG, Geertsma-Kleinekoort WM, et al. SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels. *BMC Genomics*. 2008;9:41.
11. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5(4):557-572.
12. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007;17(11):1665-1674.
13. CG (2013). Complete Genomics Diversity set repository. Retrieved 1 December 2013, from ftp://ftp2.completegenomics.com/Feb2011_Release/Diversity/.
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
15. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
16. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6(9):677-681.
17. Nimblegen (2013). SeqCap EZ Human Exome Library v3.0. from http://www.nimblegen.com/products/seqcap/ez/v3/index.html.
18. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185-193.
19. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: new insights in genome diversity. *Genome Res*. 2006;16(8):949-961.
20. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444-454.
21. Cattell RB. The Scree Test For The Number Of Factor. *Multivariate Behav Res*. 1966;1(2):245-276.
22. Lee YH, Ronemus M, Kendall J, et al. Reducing system noise in copy number data using principal components of self-self hybridizations. *Proc Natl Acad Sci U S A*. 2012;109(3):E103-110.

23. Breems DA, Van Putten WL, De Greef GE, et al. Monosomal karyotype in acute myeloid leukemia: a better indicator of poor prognosis than a complex karyotype. *J Clin Oncol*. 2008;26(29):4791-4797.

24. Mullighan CG, Su X, Zhang J, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N Engl J Med*. 2009;360(5):470-480.

25. Weischenfeldt J, Simon R, Feuerbach L, et al. Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell*. 2013;23(2):159-170.

26. Groschel S, Lugthart S, Schlenk RF, et al. High EVI1 expression predicts outcome in younger adult patients with acute myeloid leukemia and is associated with distinct cytogenetic abnormalities. *J Clin Oncol*. 2010;28(12):2101-2107.

27. Groschel S, Schlenk RF, Engelmann J, et al. Deregulated expression of EVI1 defines a poor prognostic subset of MLL-rearranged acute myeloid leukemias: a study of the German-Austrian Acute Myeloid Leukemia Study Group and the Dutch-Belgian-Swiss HOVON/SAKK Cooperative Group. *J Clin Oncol*. 2013;31(1):95-103.

28. Lugthart S, Figueroa ME, Bindels E, et al. Aberrant DNA hypermethylation signature in acute myeloid leukemia directed by EVI1. *Blood*. 2011;117(1):234-241.

29. Radtke I, Mullighan CG, Ishii M, et al. Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proc Natl Acad Sci U S A*. 2009;106(31):12944-12949.

30. Whitman SP, Liu S, Vukosavljevic T, et al. The MLL partial tandem duplication: evidence for recessive gain-of-function in acute myeloid leukemia identifies a novel patient subgroup for molecular-targeted therapy. *Blood*. 2005;106(1):345-352.

31. Dohner K, Tobis K, Ulrich R, et al. Prognostic significance of partial tandem duplications of the MLL gene in adult patients 16 to 60 years old with acute myeloid leukemia and normal cytogenetics: a study of the Acute Myeloid Leukemia Study Group Ulm. *J Clin Oncol*. 2002;20(15):3254-3261.

32. Amarasinghe KC, Li J, Halgamuge SK. CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics*. 2013;14 Suppl 2:S2.

33. Krumm N, Sudmant PH, Ko A, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res*. 2012;22(8):1525-1532.

34. Gordon DJ, Resio B, Pellman D. Causes and consequences of aneuploidy in cancer. *Nat Rev Genet*. 2012;13(3):189-203.

35. Mrozek K, Heerema NA, Bloomfield CD. Cytogenetics in acute leukemia. *Blood Rev*. 2004;18(2):115-136.

36. Cappuzzo F, Varella-Garcia M, Shigematsu H, et al. Increased HER2 gene copy number is associated with response to gefitinib therapy in epidermal growth factor receptor–positive non–small-cell lung cancer patients. *Journal of Clinical Oncology*. 2005;23(22):5007-5018.

37. Go H, Jeon YK, Park HJ, Sung S-W, Seo J-W, Chung DH. High MET gene copy number leads to shorter survival in patients with non-small cell lung cancer. *Journal of Thoracic Oncology*. 2010;5(3):305-313.

CHAPTER 12

Summary and general discussion

1. SUMMARY

The advent of array-based and next generation sequencing (NGS) technologies has increased our knowledge about the underlying genetic abnormalities in human acute myeloid leukemia (AML). Subsequently this led to further molecular categorization reflected by the recognition of AML entities defined by genetic mutations (WHO 2008). It is increasingly recognized that the epigenetic landscape, beside the genetic landscape, plays a pivotal role in leukemic transformation as evidenced by recurrent epigenetic alterations. The work presented in this thesis provides a detailed account on the use of genome-wide approaches in conjunction with novel bioinformatic methodologies to further progress the understanding of human leukemogenesis. Rapid succession in technological advances has generated a multitude of analytical tools for the genome-wide characterization of AML, with a central role for gene expression profiling (GEP), DNA mapping arrays and NGS in the work presented in this thesis.

In **chapter 2**, we devised a statistical framework utilizing the multinomial logistic regression model with a modified group lasso penalization scheme enabling sparse gene expression signature determination and the prediction of multiple classes simultaneously. A gene expression signature comprising a limited number of genes is produced for all predetermined classes and a gene within this signature is weighted according to the impact of its expression on the class prediction. We established, in agreement with previous observations, that the AML entities with favorable cytogenetics, i.e., t(15;17), t(8;21) and inv(16), can be predicted with maximum accuracy and that the gene expression signatures reflect their molecular characteristics. Additionally, we demonstrated that the framework provides a more accurate classification of AML cases harboring a combination of mutations in the genes *NPM1* and *FLT3* (*FLT3*-ITD). In **chapter 3**, we investigated the clinical outcome and gene expression signatures of *CEBPA*^{dm} and *CEBPA*^{sm} AML cases, i.e., biallelic mutations or monoallelic mutations in *CEBPA*, and their respective mutational spectra. We observed that *CEBPA*^{dm} patients are characterized by low numbers of concurrent mutations, while *CEBPA*^{sm} patients frequently harbor concurrent mutations in the genes *NPM1* and *FLT3*. We demonstrated that the *CEBPA*^{sm} marker bears no prognostic significance and, contrastingly, we reported that *CEBPA*^{dm} is a strong prognostic marker associated with favorable outcome. We employed logistic regression model with a lasso penalization scheme, reminiscent to the model described in **chapter 2** for two-class prediction, and demonstrated that *CEBPA*^{dm} AML cases can be accurately classified in a validation AML cohort while the *CEBPA*^{sm} cases do not exhibit a consistent gene expression signature.

In **chapters 4 and 5**, we investigated the genetic aberrations underlying AML and acute lymphoblastic leukemia (ALL) by utilizing DNA mapping arrays. In **chapter 4**, we reported on a software package enabling the visualization of data from DNA mapping arrays with the capacity to integrate GEP data. Copy number variation (CNV) profiles of multiple patients can be displayed simultaneously enabling the identification of recurrent genetic aberrations. In addition, we constructed a hidden markov model for the inference of loss-of-heterozygosity (LOH) regions

for AML cases without data derived from remission or normal tissue control material. In **chapter 5**, we investigated the recurrence of genetic aberrations in adult ALL and AML cases by DNA mapping arrays and NGS. We demonstrated that ALL is characterized by a multitude of recurrent small deletions and amplifications, while AML has scarce recurrent genetic aberrations. We observed genetic aberrations common in ALL or specific for B-ALL or T-ALL. Strikingly, all T-ALL cases acquired a deletion of the genes *CDKN2A/B* or genes involved in the same pathway. We discovered that the proximal genes *NF1* and *SUZ12* are commonly deleted in a subset of T-ALL and AML patients, strikingly, both genes were significant down regulated. Deep sequencing of the remaining *NF1* wild-type allele demonstrated concurrent mutations. In line with previous reports we postulated that the loss of *NF1*, activating the RAS pathway, cooperates with the loss of the polycomb repressive complex 2 (PRC2). Finally, we observed specific B-ALL cases with deletions predominantly involving promoters or the first few exons of genes. Whole exome sequencing (WES) and targeted resequencing of deleted regions in 5 B-ALL cases revealed cryptic recombination signal sequences (RSSs) flanking the breakpoints, on one or both sides, for 91% of the somatic deletions and the insertion of random nucleotides at the breakpoints. *De novo* motif detection demonstrated that the deletions are flanked by 12-bp and 23-bp spacer RSS motifs implying that the rearrangements are the result of illegitimate V(D)J recombination. Analyses of epigenetic data derived from a B-lymphoblastic cell line revealed that the breakpoints are enriched for H3K4me3, H3K27ac and RNA polymerase II binding reminiscent to antigen receptor rearrangement foci. Cross-species analysis revealed that the human breakpoints are enriched for Rag2 binding in murine thymocytes. Finally, we demonstrated that RAG-rearrangements invoke open-and-shut joints of RSS motifs likely through error-prone non-homologous end joining (NHEJ).

Chapters 6 through 10 focus on the understanding of the AML genetic and epigenetic landscape determined by NGS approaches while **chapter 11** focuses on a novel algorithm for detecting CNVs from NGS data. In **chapter 6,** we utilized WES for determining the longitudinal mutational spectrum of a severe congenital neutropenia (SCN) patient who develops AML after 17 years of G-CSF treatment. We detected in total 12 somatic alterations, including mutations in the genes *RUNX1*, *ASXL1* and *SUZ12,* in the leukemic phase and revealed that 3 mutations were already present in the early SCN phase, including the truncating mutation *CSF3R-d715*. In the leukemic phase an additional *CSFR3* mutation, i.e. *CSF3R-T595I*, is acquired, on the same allele as the truncating mutation, conferring G-CSF independence. In **chapter 7**, we investigated the underlying mechanism driving the overexpression of the proto-oncogene *EVI1* upon the acquisition of inv(3)(q21q26.2) or t(3;3)(q21;q26) (inv3/t(3;3)) in myeloid malignancies. Targeted resequencing of the breakpoint loci revealed an asymptotic pattern of breakpoint positions in the 3q21 locus, implying that a breakpoint-free common translocated segment (CTS) always repositions towards the 3q26 locus. We demonstrated by integrating RNA-Seq, Chip-Seq and 4C-Seq data that a distal *GATA2* enhancer is located within this CTS and upon repositioning

interacts with the *EVI1* promoter driving its overexpression. Concurrently, *GATA2* expression is reduced and only expressed from the non-rearranged allele. Ablation of the ectopic enhancer in the cell line MUTZ-3, by genome editing technology, demonstrated the abrogation of *EVI1* expression and subsequent growth inhibition and differentiation of the cell line. Finally, upon translocation of the 3q21 locus a super-enhancer is formed demonstrated by the high levels of H3K27ac spanning the complete CTS region. Treatment of the inv(3)/t(3;3) cell lines MOLM1 and MUTZ-3 with a BET-bromodomain inhibitor (JQ1) demonstrated abrogation of *EVI1* expression, growth inhibition, differentiation and increased apoptosis reminiscent to the genome editing experiment, while this was not observed in the *EVI1* overexpressing cell line K562. In **chapter 8**, we determined the mutational spectrum of inv(3)/t(3;3) myeloid malignancies by WES and RNA-Seq. We observed that 98% of the inv(3)/t(3;3) cases acquire activating mutations in the RAS/RTK signaling pathways. In addition, mutations in the remaining wild type allele of *GATA2* as well as heterozygous alterations in *SF3B1*, *RUNX1* and genes encoding epigenetic modifiers frequently co-occured with the inv(3)/t(3;3) aberration. Notably, we observed neither differences in mutational patterns nor GEP across inv(3)/t(3;3) myelodysplastic syndrome (MDS), chronic myeloid leukemia in blast crisis (CML-BP) and AML cases, suggesting that inv(3)/t(3;3) myeloid malignancies could be recognized as a single disease entity. In **chapter 9,** we employed RNA-Seq on leukemic blasts from a patient which progressed from MDS to AML. We discovered that the unreported *KMT2A-MYH11* fusion transcript was already present in the MDS phase. In **chapter 10**, we investigated AML cases with an amplification involving the gene *BCL11B* and its consequent overexpression. Targeted resequencing of the *BCL11B* locus in these AML cases revealed that the additional copy is the consequence of a jumping translocation by which the donor *BCL11B* locus integrates into super-enhancer regions located on different chromosomes. In **chapter 11**, we provided a novel statistical framework for estimating CNV profiles from NGS data. We constructed an algorithm that infers systematic noise from a reference dataset comprising diploid cases, i.e., healthy individuals or remission material, and demonstrated that these inferred systematic noise components can be utilized to generate accurate CNV profiles from whole genome sequencing (WGS) and WES data. In comparison to classical DNA mapping arrays, we demonstrated that the framework detects more CNVs corroborated by structural variants.

2. GENERAL DISCUSSION

This final part puts the results and remaining hypotheses, provided in this thesis, in the perspective of contemporary knowledge and their possible implications are discussed in detail. The work presented in this thesis was divided in three parts each addressing different research questions by employing a specific genome-wide approach. Concluding, the prospectives and perspectives of future leukemic research are provided with a focus on mutational and functional relevance, clonal evolution, genome editing and translational medicine.

2.1 Gene expression profiling in acute myeloid leukemia

Initially, transcriptome-driven analyses revealed the heterogeneity of AML and the existence of molecularly defined AML subtypes.[1-3] It remained a pending question whether GEP-based classification could replace traditional diagnostic methodologies, e.g., cytogenetics and standard PCR-based techniques, or would only serve as a discovery tool for research purposes. Therefore, additional experiments were conducted demonstrating that GEP profiling is valuable for classifying AML entities with favorable cytogenetics[2], i.e., characterized by t(15;17), t(8;21), inv(16) cytogenetic abnormalities, and AML cases with mutations in the genes *CEBPA*[4] and *NPM1*[5], however, predictive signatures could not be found for AML cases with mutations in RAS or receptor tyrosine kinase (RTK) affiliated genes, e.g., *FLT3*, *NRAS* and *KRAS*. Reasons for the miss-classification of AML cases were focused on the greater than expected heterogeneity within AML subgroups, the number of AML cases, the minimal effect of activating RAS/RTK mutations on gene expression levels and the classification algorithm utilized. The absence of predictive value for a molecular AML marker could be due to the following: (I) the effect of the genetic alteration on gene expression levels is minimal, (II) the gene expression signature conferred by a genetic alteration is overpowered by that of a concurrent mutation. The first situation is intractable and could necessitate the use of different genome-wide approaches for class prediction purposes, e.g., detection of epigenetic alterations. The second situation necessitates statistical frameworks enabling the modeling of concurrent genetic alterations thereby generating independent gene expression signatures. In **chapter 2** we devised a multinomial logistic regression model with a modified group lasso penalization scheme[6,7] enabling multiple class prediction and providing weighted sparse prediction signatures. Prediction signatures were established from a training AML cohort and validated on a large independent AML cohort for the following situations: (I) AML cases with mutually exclusive favorable cytogenetic abnormalities reflecting transcriptional homogeneous AML subgroups, (II) AML cases harboring mutations in the gene *NPM1* (*NPM1*$^+$) and/or internal tandem duplications in the gene *FLT3* (*FLT3-ITD*) reflecting transcriptional heterogeneous AML subgroups. The first experiment demonstrated that these AML entities are predicted with maximum accuracy, while the second experiment demonstrated increased prediction accuracy in comparison to a previous study[2], attributed to the adopted multiple class prediction procedure. In the previous study, the molecular heterogeneity was not captured due to the two-class prediction procedure employed as the *NPM1*$^+$/*FLT3-ITD* specific expression pattern confounded the predictive signatures of the *NPM1*$^+$ or *FLT3-ITD* markers alone. In molecular heterogeneous situations GEP-based classification has debatable additive value with respect to standard PCR detection techniques, however, for research purposes could be utilized for novel class prediction or the provision of gene expression signatures reflecting pivotal biological processes underlying the disease of interest.

Different studies utilized gene expression signatures for prognostication.[8,9] Gene expression signatures derived from functionally validated leukemic stem cells (LSCs) associate with inferior

clinical outcome in AML.[10-12] This association reflects the persistence of stemness or stem cell programs in the leukemic blasts in poor outcome AML subtypes. Prognostication, likewise to GEP-based classification, is limited by molecular heterogeneity precluding the identification of prognostically valuable gene expression markers. Correlation of gene expression patterns presents another conundrum affecting discrimination and prognostication of AML subtypes. An important consideration is whether novel predictive gene expression markers add prognostic value to already established prognostic parameters or are completely redundant. Previously, we demonstrated that *MLL*-rearranged AML cases are stratified into a prognostically favorable and intermediate group based on gene expression levels of the gene *BRE*, demarcating a *MLL*-rearranged subgroup with a strong gene expression signature.[13] Previous reports demonstrated that *EVI1* gene expression levels enable the stratification of *MLL*-rearranged AML cases in a prognostically poor and intermediate group[14] and is anti-correlated to *BRE*[15], rendering the question which gene expression marker is prognostically more accurate. Efforts should be directed towards the development of statistical frameworks accounting for molecular heterogeneity and gene expression correlation structures for classification and prognostication purposes. In conclusion, GEP enables the accurate classification and prognostication of a limited set of AML entities, however, molecular heterogeneity precludes further classification unless further subcategorizations can be introduced.

2.2 Copy number variations

DNA copy number variations (CNVs) are genetic hallmarks of cancer.[16,17] The advent of DNA mapping arrays enabled the high-resolution identification of CNVs, however, the particular challenge is to identify CNVs perturbing specific cancer-related genes. Most CNVs are large enough to affect multiple genes thereby invoking the "passenger-driver" obstacle, implying that particular genes drive oncogenesis while others are affected by mere coincidence.

2.2.1 Copy number variations in acute leukemia: recurrence and origin

Genetic lesion recurrence is valuable for further subcategorization of molecular subtypes. In **chapter 5**, we determined genetic alterations in 173 adult acute leukemia cases by DNA mapping arrays. Strikingly, all T-ALL cases acquired CNVs perturbing the *CDKN2A/B* pathway. The deletion of *CDKN2A* disrupts cell cycle control and frequently co-occurs with *NOTCH1* mutations which affects the self-renewal capacity of cells.[18,19] The relevance of co-occurrence and the underlying mechanism of the frequent deletion of *CDKN2A* remains unanswered and a topic for future research.

We observed a deletion affecting the proximal genes *NF1* and *SUZ12* in 5 AML and 3 T-ALL cases. NF1 loss activates the RAS pathway due to its inhibitory function[20], while SUZ12 is a pivotal member of the PRC2 complex.[21] Subsequent GEP analysis revealed that both genes were significantly down regulated, giving interesting perspectives as recent reports demonstrated

that loss of NF1 cooperates with loss of PRC2 in malignant peripheral nerve sheath tumors and melanoma.[22,23] Deep targeted resequencing of the remaining *NF1* wild type allele revealed premature stop codon introducing mutations imparting its complete loss. Following the same line of reasoning as the previous reports, we expect that RAS activation cooperates with PRC2 loss in a subset of patients with AML or T-ALL. It is valuable to investigate if the reduced *SUZ12* expression causes a disintegration of the PRC2 complex, subsequently reducing H3K27me3 levels and changes in gene expression levels.

Previous studies demonstrated that pediatric and adult ALL acquire recurrent CNVs affecting genes involved in lymphoid development.[24-26] A recent study demonstrated that the recurrent genetic lesions are partially invoked by RAG-mediated rearrangements.[27] Targeted resequencing of the deletions in 5 *BCR-ABL1*/*BCR-ABL1*-like cases revealed that the breakpoints were predominantly flanked (91%) by cryptic RSS motifs, implying RAG-mediated rearrangements in almost all deletion events. A pending question remains why the illegitimate RAG-mediated rearrangements occur outside the antigen receptors in ALL. Fulfillment of these rearrangements must meet a few prerequisites: (I) expression of the proteins RAG1 and RAG2, (II) presence of 12/23-bp spacer RSS motifs, (III) accessible DNA marked by H3K4me3 and H3 acetylation, (IV) CTCF binding and long range interactions, (V) a functional NHEJ pathway.[28-30] Most of the B-ALL cases continually express *RAG1* and *RAG2*, or temporarily expressed these proteins as evidenced by extensive V(D)J

rearrangements. We provided evidence that the breakpoints are predominantly flanked by cryptic RSS motifs and enriched for active chromatin markers. In all likelihood the long range interactions occurred due to the proximity of the breakpoints during repair. The increased activity of illegitimate RAG-mediated rearrangements could be twofold: (I) increased activity of the RAG complex, (II) deficiency in the DNA repair capacity. The HMG-box family proteins HMGB1 and HMGB2 are required for RAG complex assembly and modulate its activity by bending the DNA in a catalytically favorable manner for cleavage.[31,32] The increased activity hypothesis provides an interesting perspective as the RAG2 protein harbors a PHD domain enabling the binding to H3K4me3.[29,33] Rag2 Chip-Seq data derived from murine thymocytes revealed the genome-wide binding to H3K4me3 enriched loci[33] deposited mainly in promoter and enhancer regions predisposing these regions to RAG-mediated rearrangements. We did not find any changes in *HMGB1* or *HMGB2* expression levels and efforts should be directed to infer if there are changes on the protein level or through posttranslational modifications. The deficient DNA repair pathway hypothesis provides another interesting perspective as the sequenced cases comprised *BCR-ABL1* and *BCR-ABL1*-like cases. Previous studies demonstrated that c-ABL and BCR-ABL1 interact with and down regulate the pivotal NHEJ DNA repair protein DNA-PKcs.[34,35] This warrants the study of DNA-PKcs protein expression level or its phosphorylation by *BCR-ABL1* or other kinase-activating lesions. Whether the RAG activation or the deficient repair pathways confers illegitimate RAG-mediated rearrangements remains yet elusive.

2.2.2 Driver or passenger: one car multiple seats

CNVs predominantly affect multiple genes simultaneously and determining the specific gene driving oncogenesis remains difficult. Different studies resort to defining a minimally affected region (MAR), however, this procedure would preclude the identification of cancer-related genes cooperatively affected or those flanking the recurrent CNV. For example, in **chapter 5**, we identified the recurrent deletion simultaneously affecting the promoter of the gene *MKKS* and the gene *SLX4IP* (Figure 1A). We observed that *SLX4IP* expression levels remain unperturbed while *MKKS* expression levels became significantly reduced (Figure 1B). The MAR procedure would have identified *SLX4IP* as a potential oncogenic driver, while GEP would have highlighted *MKKS* as a potential oncogenic driver. Promoter deletion is common in B-ALL pathogenesis and warrants a different cancer-related gene detection approach. We adopted a kernel density procedure with a flat top normal distribution kernel[36] for the identification of recurrently affected genes. This procedure enriches for genes affected by genetic lesions if it is: (I) frequently affected, (II) specifically affected or (III) frequently flanking breakpoint loci. The disadvantage of this methodology is that it still relies on frequencies for enriching genes.

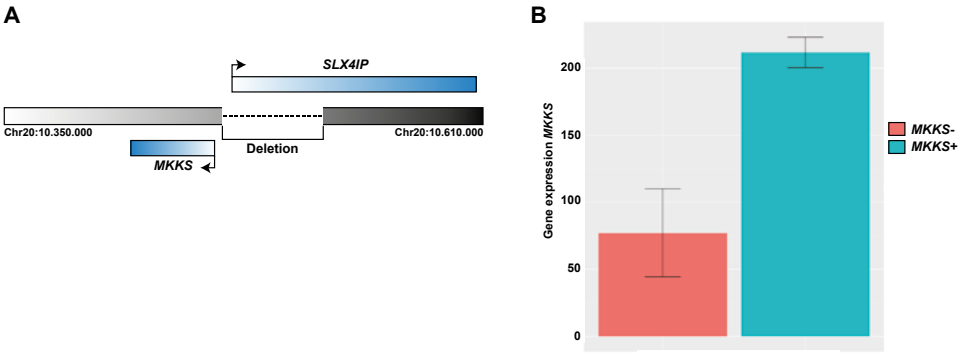


Figure 1. Deletion perturbs gene expression levels *MKKS*. (**A**) Common deletion of the *MKKS* promoter in ALL, (**B**) *MKKS* is significantly down regulated in ALL cases with the deletion of the *MKKS* promoter (*MKKS-*, $p < 0.001$).

2.3 Next generation sequencing

2.3.1 Next generation sequencing in AML

Whole exome and genome sequencing

The recent introduction of next generation sequencing (NGS) commenced the determination of the genetic and epigenetic landscape of different human diseases.[37-39] The broad spectrum of NGS approaches enables the molecular characterization of cancer material. Chief amongst those approaches are whole genome sequencing (WGS) and whole exome sequencing (WES)

enabling the determination of the mutational spectrum of cancer material, e.g. *de novo* AML.[40] The mutational spectrum provides details related to the process of leukemogenesis, concurrence or mutual exclusivity of mutations, and signatures of mutational processes.[41] In **chapter 6**, we determined the longitudinal mutational spectrum of a SCN patient progressing towards AML. Specific focus is put on the process of leukemic transformation and clonal evolution, however, a large gap of 9 years between the SCN and leukemic phase precludes the accurate identification of mutations driving leukemic transformation as some might be acquired during the pre-leukemic phase or lost due to clonal tides.[42] In **chapter 8**, we conducted WES on inv(3)/t(3;3) myeloid malignancies and demonstrated mutations affecting the RAS/RTK signaling pathways, *GATA2*, *SF3B1*, *RUNX1* and epigenetic modifiers. Both chapters use WES for determining the mutational spectrum and the process of clonal evolution. Novel tools are continuously developed for the analysis of WGS and WES data. Variant detection analysis tools, such as SAMtools[43] or MuTect[44], sometimes sub-optimally analyze the data. The different variant detection lists generally overlap, however, there are some differences. In the end, the final variant list is created by a pipeline which efficiently combines all variant lists produced by different variant detection tools. Future efforts should focus on developing a variant detection tool capable of accounting for all requisites established in the various variant detection tools.

Alignment algorithms accidentally introduce false positive mutations, especially in regions containing repeat elements or multiple complex SNPs. Recent efforts produced variant detection algorithms based on *de novo* assembly procedures[45] which could, with a proper germline control, resolve this issue. Standard alignment algorithms commonly introduce insertions and deletions in reads to achieve genome alignment, frequently resulting in ambiguous alignments. *De novo* assembly procedures disregard the target genome and produce linear assembled sequences, making them directly comparable between case and control.

CNV profiles can be determined from WGS or WES data[46,47], yet many algorithms are impaired by the systematic noise inherent to the sequencing technology. In **chapter 11**, we constructed a novel statistical framework for determining CNVs from WGS and WES data by taking into account systematic noise extracted from a diploid reference set. This methodology attenuated systematic noise and accurately detects more CNVs than DNA mapping arrays. Additionally, the framework is applicable to NGS data derived from different organisms (data not shown). This study implies that traditional DNA mapping arrays could be replaced by NGS derived CNV estimates due to its increased accuracy and detection resolution.

Whole transcriptome sequencing

Transcription deregulation plays an important role in leukemic transformation. RNA-Seq enables gene expression level estimation, transcript isoform detection, transcript fusion detection, and variant detection. The pending question remains if RNA-Seq can completely replace microarray technology, traditional cytogenetics and PCR-based techniques, for determining

molecular AML entities and prognostication.[48] RNA-Seq, in addition to the gene expression level estimation, enables the determination of recurrent cytogenetic abnormalities by fusion transcript detection[49], including information of the fusion transcript exon structure making it comparable to standard PCR detection. Interestingly, fusion detection in RNA-Seq supersedes standard PCR detection in some aspects as it enables: (I) the detection of multiple fusion transcripts in one experiment, (II) the detection of novel, complex or non-canonical fusion transcripts. RNA-Seq, in contrast to microarray technology, is not limited to a predefined set of measurable loci and enables the detection of more transcripts including long non-coding transcripts. Gradually, traditional microarray technology and standard PCR detection protocols will be replaced by RNA-Seq approaches due to its generality and the additional molecular information provided.

Epigenetics

Epigenetic alterations represent an additional layer of control conferring regulation of cellular processes. These alterations, e.g., DNA methylation or histone methylation, are dynamically removed or deposited.[50] Epigenetic characterization of leukemic blasts by array-based approaches, e.g., tilling and Chip-on-Chip arrays, demonstrated that AML is also heterogeneous at the epigenetic level.[51] Additionally, mutational profiling demonstrated that AML is characterized by frequent mutations in epigenetic modifiers.[52,53] Recent studies examining the epigenetic landscape of cancer demonstrated the existence of super-enhancers driving the expression of proto-oncogenes, e.g. *MYC*.[54]

NGS approaches enable the determination of many different aspects of the epigenetic landscape in detail: (I) DNA methylation: bisulphite sequencing, MeDIP-Seq, (II) Protein binding and histone alterations: Chip-Seq, (III) chromatin conformation: 4c-Seq, HiC, ChiA-PET, (IV) open chromatin: DNAse-Seq. The many different NGS approaches developed and the increased resolution it provides implies the gradual replacement of array-based technologies.

Epigenetic regulation is a dynamic process mandating a demarcated experiment for optimal detection, e.g, type of epigenetic alteration, cell type or timing, as determining these alterations *ad libitum* could present confounding results. The unprecedented epigenetic characterization of AML enables the determination of dynamic leukemogenic mechanisms, however, it will also likely introduce additional sources of heterogeneity complicating AML subcategorization.

2.3.2 Integrative approaches

Determining the mutational, transcriptional or epigenetic landscape enables further delineation of the leukemogenic process. Further improvements in understanding leukemogenic processes should be expected from integrating these data types. Epigenetic alterations modulate the transcriptional activity of target genes, while mutations modulate genome-wide epigenetic patterns, implying interactions on different levels. In **chapter 7**, we demonstrated that the proto-oncogene *EVI1* is overexpressed due to the repositioning of a distal *GATA2* enhancer through

the integration of targeted resequencing, RNA-Seq, Chip-Seq and 4c-Seq data. In **chapter 10**, we demonstrated that the gene *BCL11B* becomes overexpressed due to an amplification of the encompassing region as detected by DNA mapping arrays and GEP. Subsequent breakpoint detection by targeted resequencing revealed that a jumping translocation involving *BCL11B* integrates into loci characterized by super-enhancer elements driving the overexpression of the translocated *BCL11B*. These chapters provide examples of integrative approaches utilized for understanding the underlying mechanistic process associated with leukemogenesis.

Continuous efforts should be put into the development of novel statistical frameworks for the analysis and integration of NGS data. The introduction of public NGS databases and projects, e.g., the cancer genome atlas (TCGA), requires researchers to distribute their published data and enables procurement of already generated or published data. The number of public data sets is rapidly increasing and renders generating novel data superfluous, except for very specific research questions. Developments in the fields of bioinformatics and statistics introduced novel frameworks for integrating NGS datasets. Some of these data integration tools are very specific, e.g., the detection of mono-allelic expression[55,56], while others are more general[57] and are used for integrating different types of NGS data. Although general data integration tools are beneficial, most specific research questions are addressed by specialized integration methodologies. Additionally, novel NGS approaches are continuously developed which require specific data processing techniques. The development of data integration tools is an emerging field and, given the limited number of tools, requires international efforts for bringing it to maturity.

2.4 Prospects and perspectives on future leukemia research

2.4.1 Basic understanding of mutation functionality and disease development

Functional consequences of mutations

In recent years we have witnessed an increase of NGS studies identifying novel genetic lesions associated with cancer development. AML is one of the most well characterized cancer types and, in all likelihood, the number of newly detected genetic lesions will decrease rapidly.[40,53,58,59] Currently, the larger question at hand relates to the implications of the already discovered genetic alterations. NGS has furnished leukemic research with a wealth of additional genetic markers, however, the basic understanding of the implication of most recurrent mutations is still lacking. In addition, a considerable number of genetic lesions display patterns of co-occurrence or mutual exclusivity implying relatedness, however, a larger number of mutations are patient-specific precluding the understanding of their functional implications. Hence, the heterogeneity of AML is more than expected based on genome-wide approaches previously employed.

Enigmatic functional consequences: an example

For example, recent efforts demonstrated that *DNMT3A* is mutated in approximately 20% of the AML cases[58,60] and mutations are predominantly located in the methyltransferase domain.[61] The

gene *DNMT3A* encodes a *de novo* methyltransferase important for establishing *de novo* genomic DNA methylation patterns.[62] Given its function as a *de novo* methyltransferase, deductive thinking provided the hypothesis that mutations impair the establishment of genome-wide DNA methylation patterns thereby driving leukemic transformation. Initial experiments demonstrated minimal changes in genome-wide methylation levels[58,60] and subsequent studies observed only limited and focal DNA methylation changes[63-65], contrastingly, at some loci increased methylation levels. The correlation between methylation level differences and gene expression levels is considered weaker than expected in *DNMT3A* mutated AML cases, especially when corrected for the frequently concurrent *NPM1* mutation.[58,60,63] The question remains why mutations in the methyltransferase domain have such limited effect. Recent deep sequencing studies demonstrated that *DNMT3A* mutations are sometimes pre-leukemic[66,67] and detectable in healthy individuals.[66] These pre-leukemic clones can take years before they progress towards leukemia.[68] The pending question remains if *DNMT3A* mutations truly invoke leukemic conditions by altering DNA methylation levels as: (I) a substantial number of mutations are observed outside the methyltransferase domain, (II) methylation changes are limited and focal, (III) mutations are found in pre-leukemic clones, taking years to develop, and healthy individuals. Leukemic progression in these cases could be due: (I) a necessary concurrent mutation in another gene is acquired resulting in synergistic effects driving leukemic development, (II) mutations in *DNMT3A* predisposes for leukemia by another currently unknown mechanism.

Previous studies linked *DNMT3A* aberrations in mouse embryonic stem cells (mESC), including ablation, to DNA hypomethylation[64,65,69], subsequently associating it to AML. Recent studies in the overgrowth syndrome, a group of disorders characterized by an abnormal increase of the body or a body part, demonstrated congenital *DNMT3A* mutations previously identified in AML.[70] Although an increased incidence of cancer has been observed in overgrowth syndrome subgroups, these patients do not specifically develop AML. These two studies provide contrasting accounts concerning *DNMT3A* mutations demonstrating that their functional consequences still remain enigmatic.

Basic understanding of mutations in leukemogenesis

Overall, the assessment of mutational functionality is a difficult task with many pitfalls. The mutations in *DNMT3A* remain functionally enigmatic, however, this holds for many genes involved in leukemic transformation, e.g., *NPM1*. The basic understanding of gene mutations is fundamental for progressing the mechanistic knowledge of leukemogenesis and becomes even more complex when considering concurrently mutated genes. Initially, novel gene mutations are postulated to affect the primary function of the protein it encodes, although logical, this might not always provide the best hypothesis. For instance, *IDH1* and *IDH2* mutations result in a neomorphic function[71] and *NPM1* mutations introduce a nuclear export signal[72], both partially or completely unrelated to their primary protein function. The advent of NGS resulted in the detection of many

novel genetic lesions, however, what these lesions functionally engender remains a topic of future leukemic research.

2.4.2 Leukemic heterogeneity: making sense of mutational patterns

Mutation patterns, i.e. concurrence or mutual exclusivity of genetic lesions, provide an avenue for understanding how recurrent genetic lesions contribute to leukemic development. Genes more frequently mutated can be assumed to have a greater impact on disease development. Furthermore, mutation patterns are better discernable from genes frequently mutated. Mutational patterns exhibiting mutual exclusivity hint towards genes with similar functions. For example, aggregating gene mutations into pathways demonstrates how frequently a particular process is affected. In **chapter 8**, we observed that the RAS/RTK signaling pathway is affected in 98% of inv(3)/t(3;3) myeloid malignancies.

For example, mutational pattern analysis revealed that mutations in the genes *TET2*[73], *IDH1* and *IDH2* are mutually exclusive implying functional relatedness. Recent efforts demonstrated that all mutations within these genes impair the conversion of 5-methylcytosine to 5-hydroxymethylcytosine resulting in genome-wide hypermethylation.[74,75] Patterns of mutual exclusivity were also observed for mutations in the splicing machinery[76] and cohesion complex genes.[53]

Information concerning synergistic effects can be inferred by ascertaining patterns of concurrence. Mutations in *DNMT3A*, *NPM1* and internal tandem duplications in *FLT3* are frequently co-occurring, however, the underlying synergistic mechanism remains yet to be elucidated.

The strength of discerning mutational patterns is highly dependent on: (I) how frequently the genes are mutated, (II) frequency of concurrence, (III) frequency of exclusivity, or (IV) the total number of patients screened. Genetic alterations observed in a single or a few AML cases remain difficult to interpret and could only be investigated by experimental approaches or put in the context of contemporary knowledge. The integration of different data types, e.g., RNA-Seq or Chip-Seq, could further help elucidate the underlying mechanisms as demonstrated in **chapter 7**.

2.4.3 Clonal evolution

NGS approaches detect variants at different allele frequencies, therefore enabling the determination of the clonal composition.[77] Recent time-series or diagnosis-remission-relapse trios studies demonstrated the existence of oligoclonality and variegated clonal evolution in AML.[78,79] Small subclones, whether or not derived from the dominant clone, can have survival advantages. Therapeutic abrogation of the dominant clone enables the treatment insensitive subclone to expand and confer leukemic relapse. In therapy-related AML these subclones could have already been present, without developing leukemia, before therapy induction.[80] In addition, the subclone can acquire additional mutations making it potentially more malignant. Clinically, these subclones remained undervalued due to the detection limitations of cytogenetics and standard

PCR techniques. Studies investigating the clonal composition of AML demonstrated strong clonal evolution propensities, precluding the persistence of remission, in specific AML subgroups.[78] For example, mutations in *DNMT3A*, *TET2*, *IDH1* and *IDH2* have been demonstrated to exist in pre-leukemic clones.[66,81] Pre-leukemic clones harboring *DNMT3A* mutations remain unperturbed by current treatment protocols as they remain detectable in remission material and confer a strong relapse risk, strikingly, sometimes after several years.[66] Discerning the clonal architecture and understanding the dynamics of clonal evolution provides therapeutic actionable options and insight into relapse initiating processes. Treatment modalities can be modulated or combined with the knowledge of the complete clonal architecture preventing outgrowth of subclones insensitive for standard treatment protocols.

Recent efforts focused on understanding the clonal architecture and its preceding evolution by determining which mutations are acquired in the same clone. The production of subclones occur by the following principals (Figure 2A): (I) branching: a mutagenesis sensitive pre-leukemic or leukemic clone produces multiple offspring subclones by continual acquisition of mutations, (II) linear: a pre-leukemic or leukemic clone acquires an additional mutation and produces a subclone with potentially more malignant or treatment-insensitive properties. The principles of branching and linear clonal evolution result in a clonal composition decomposable by their unique features of inheritance. However, the complete determination of the clonal dynamics is frustrated by clonal evolution caveats (Figure 2B): (I) recurrence: multiple subclones acquire independently the same subset of mutations invalidating the pattern of unique inheritance, (II) clonal tides[42]: subclones are produced by receding parental clones precluding the correct parental clone attribution. The problem of recurrence relates to clonal identifiability and correlates to the total number of clones present and genetic markers detected. The addition of detected genetic markers increases the identifiability of clonal dynamics as it improves the separation of clonal constituents. The caveat of clonal tides remains intractable, unless: (I) the missing clone can still be detected with very deep sequencing, (II) strong assumptions are made about mutation acquisition order, (III) determination of clonal composition at multiple time points enabling enhanced disentanglement of the clonal architecture.

Current efforts are directed at determining the mutual composition of each clone and the complete clonal architecture. Detecting which mutations are acquired in the same clone remains difficult. Initially, the variant allele frequency was used for inferring the clonal architecture and necessitated strong mutational pattern assumptions. For instance, mutations observed in 50% of the reads could be heterozygous in all leukemic blasts or homozygous in only half of the leukemic blasts, e.g., *SF3B1* mutations are heterozygous therefore observed in all leukemic blasts. These assumptions are produced for a limited number of well-characterized mutations, however, are rendered invalid when these mutations are acquired in (multiple) subclones. Recent statistical developments provided statistical inference algorithms for determining the clonal architecture based on very deep sequencing.[82,83] Although substantially more efficient than pre-existing

methodologies, it still necessitates *a priori* knowledge concerning: (I) copy number of the affected region, (II) LOH status of the affected region, (III) probability that the mutation is heterozygous or homozygous in the subclone. In addition, it necessitates very deep sequencing (> 1000x coverage depth) and time-series, e.g., diagnosis, relapse and second relapse samples, to determine the clonal architecture. These algorithms are rendered invalid when the principle of recurrence is established (*vide supra*) to occur.

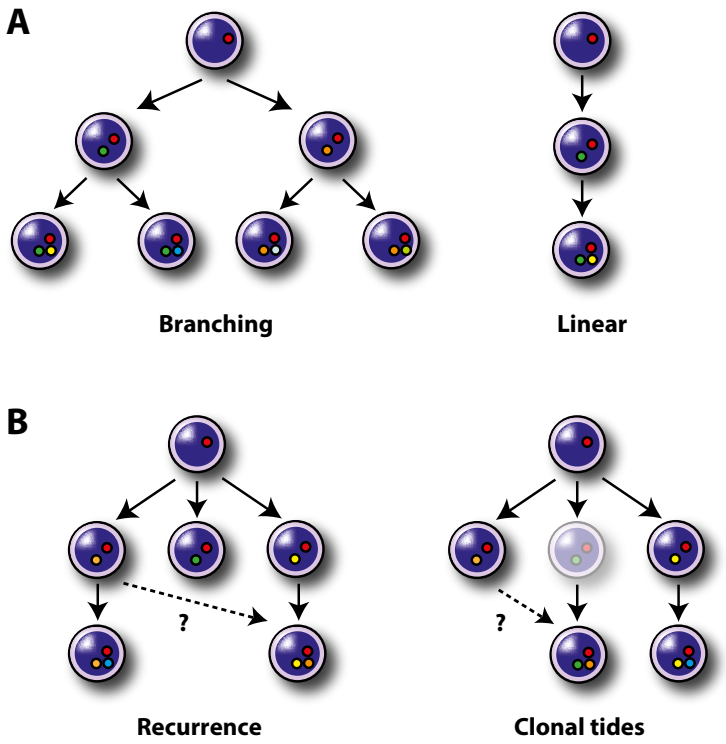


Figure 2. Mechanisms of clonal evolution. (A) Principles on clonal evolution. (Branching) A parental clone acquires mutations producing one or more subclones. Subsequently these subclones can acquire mutations and produce one or more subclones. Unique inheritance of mutations enables the delineation of the clonal architecture, (Linear) A parental clone acquires mutations and its offspring subsequently acquires additional mutations. Unique inheritance of mutations enables the determination of the clonal architecture. (B) Pitfalls in determining clonal evolution (Recurrence) Subclones are produced in a branching or linear manner, however, a mutagenic environment enables the recurrence of particular mutations in two or more subclones. Very similar clones, only differing in one mutation, can produce subclones from which the inheritance is undiscernible precluding the delineation of the clonal architecture, (Clonal tides) Like recurrence a particular mutations has been recurrently acquired except the parental clone has receded. Statistical models would attribute this subclone to the false parental clone. (Solid arrow) Correct attribution of inheritance, (Dashed arrow) incorrect attribution of inheritance.

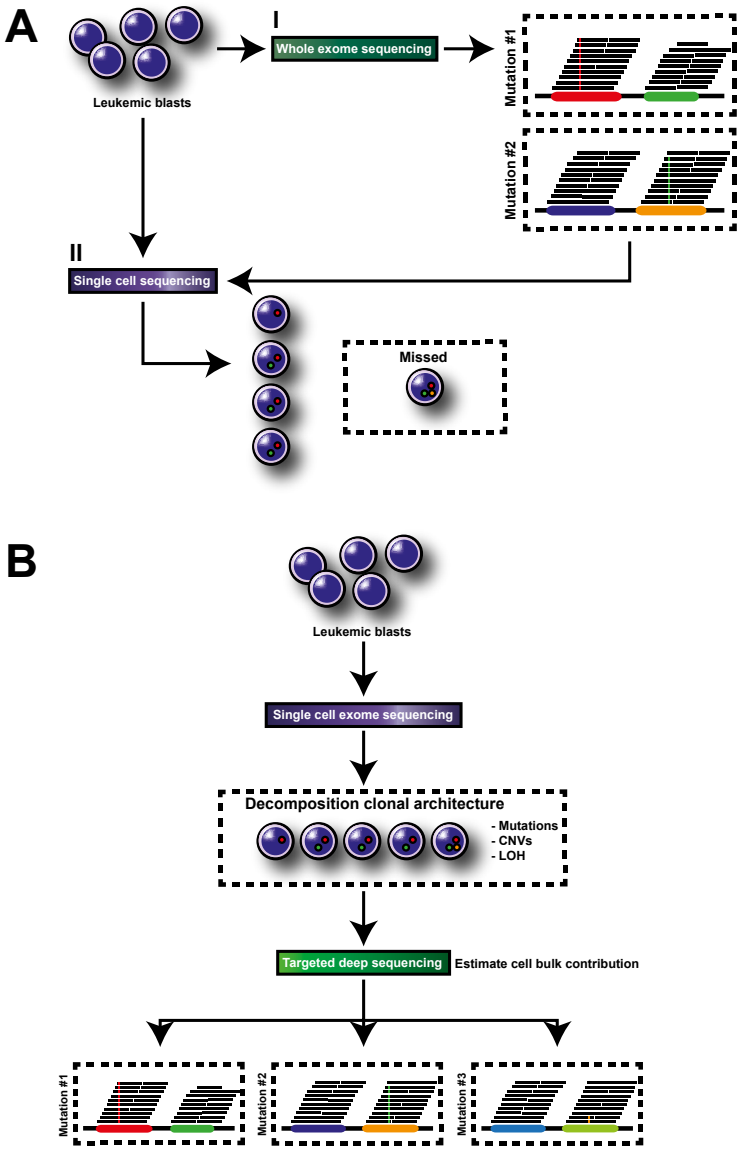


Figure 3. Decomposition of the clonal architecture. (A) (I) Genetic alterations are determined from the leukemic blasts by WES. (II) Subsequent targeted single cell resequencing of the mutations reveal the clonal architecture. Mutations missed by WES, due to reduced coverage or a very minor clone, will not be targeted and prohibit the complete decomposition of the clonal architecture. (B) WES single cell mutational analysis provides information about concurrence and mutual exclusivity of mutations, while coverage statistics and genotype information provide estimations for copy number variations and regions of loss-of-heterozygosity. Combined, these data enable the complete decomposition of the clonal architecture. Estimation of the clonal contribution to the bulk of the leukemic blasts should subsequently be estimated by targeted deep sequencing of the observed mutations.

Targeted single cell sequencing enables a more precise definition of the complete clonal architecture[84] and the introduction of single cell WES enables greater insight in clonal dynamics and architecture.[85] Two schemes are proposed for delineating the clonal architecture by combining WES/WGS with single cell sequencing. The first scheme proposes the use of WES for determining mutations in the bulk of leukemic blasts (Figure 3A(I)). These mutations are used for targeted single cell sequencing (Figure 3A(II)) enabling the determination of mutational patterns and subsequently the delineation of the clonal architecture. WES mutation detection has some disadvantages as mutations could be missed due to: (I) minimal or no coverage of the target region, (II) undiscovered low abundant clones, (III) the introduction of proximal false positive mutations by alignment algorithms.

The second scheme involves WES of multiple single cells for determining the clonal composition (Figure 3B). The total contribution of clones is difficult to estimate from single cell WES and subsequent targeted deep sequencing of the mutations in bulk material enables estimating the contribution of the clones. The second scheme produces more practical information, however, the total cost would render it impractical for analyzing a large AML cohort. The first scheme enables the detection of almost all mutations and subsequent targeted single cell sequencing enables the characterization of thousands of cells simultaneously. Finally, recently single cell approaches for RNA-Seq[86,87] and genome-wide bisulfite sequencing[88] have been adopted and enable the inference of transcriptional heterogeneity and clonal epigenetic variety. Single cell RNA-Seq enables clonal decomposition by identifying the genetic lesions in the RNA-Seq reads, reminiscent to **chapter 8**.

2.4.4 Genome editing

Genome editing is a novel field enabling the modulation of the DNA sequence or gene expression *in vivo*. Multiple proteins are able to bind specific DNA sequences and introduce DNA double strand breaks (DSB): (I) zinc finger nucleases[89], (II) transcription activator-like effector nucleases (TALENs)[90], and (III) the CRISPR/Cas9 system.[91] The induction of proximal DSBs enables the deletion of complete regions in cell lines or mouse models. Additionally, target regions can be edited by providing an exogenous DNA template during the homology directed repair (HDR) of DSBs, enabling the introduction or editing of disease causing mutations.[92] In **chapter 7**, we employed TALENs and the CRISPR/Cas9 system in the cell line MUTZ-3 to ablate the repositioned enhancer. These genome editing approaches enable the introduction of artificial chromosomal rearrangements[93] thereby modeling known translocations in an isogenic system. A recent study described a protocol for gene editing the components of the human hematopoietic stem and progenitor cell (HSPC) compartment retaining engraftment and repopulation capacity.[94] Minor changes to genome editing approaches enable the specific silencing of genes (CRISPRi) or activation of genes (CRISPRa) with minimal off target effects.[95]

Efforts should be directed towards developing genome editing approaches for understanding basic mutation functionality. In addition, it enables the modeling of concurrent genetic lesions in isogenic systems. The paradigm of clonal evolution can be modeled by observing the behavior of successfully edited and engrafted cells at steps of evolution. The isogenicity of the edited model system prevents the variability observed in: (I) murine disease models due to different in genetic backgrounds, (II) cancer cell lines being notoriously genetically unstable, (III) patient material due to additional concurrent mutations, (IV) patients overall due to different genetic backgrounds. Genome editing enables the modeling of infrequently observed mutations or subclones found with low leukemic burden contribution. Genome editing in combination with xenograft models enable further behavioral understanding of leukemic development and the effect of recurrent mutations. Determining which mutations are causally related to leukemogenesis and their functional etiology remains a major challenge, therefore modeling of mutational patterns observed in leukemia through genome editing enables the etiological delineation of leukemogenesis.[96]

2.4.5 Translational medicine

The increasing insight of the functional implications of genetic lesions should at one point translate to treatment improvements. The current treatment armamentarium of particular AML entities remains scarce, e.g., inv(3)/t(3;3) myeloid malignancies. We demonstrated that these AML cases are sensitive for BET-bromodomain inhibitors[97] and demonstrated that almost all cases acquire activating RAS/RTK signaling pathway mutations. Although many of the RAS pathway constituents are notoriously difficult to target[98,99] it provides valuable information for treatment design if they ever become targetable, e.g., *FLT3*-ITD.[100] Recent efforts led to the development of treatment modalities against epigenetic modifiers. Azacytidine prevents hypermethylation and is used for high-risk MDS treatment.[101,102] Recently, specific mutant IDH2 inhibitors have been shown to induce differentiation in primary human AML cells[103], whereas mutant IDH1 inhibitors induced expression of gliogenic genes in glioma.[104] Mislocated enzyme activity of DOTL1 is postulated as the oncogenic driver in mixed lineage leukemia and inhibition of DOTL1 results in the apoptosis of cells carrying the *MLL*-rearrangement.[105] Initial results of these studies are promising, however, if targeted therapeutic agents will ever fully replace current therapeutic protocols remains yet to be determined.

Genes or pathways investigated for treatment development should be selected on: (I) the frequency of mutations, (II) the dependency of the cell on the presence of the mutations (Achilles heel), (III) the degree of being targetable. High-throughput drug screening with a valuable output statistic, e.g., effect on proliferation, survival or resumed differentiation, enables the discovery of novel therapeutic compounds affecting AML cells with particular mutational compositions. Reverse engineering the etiology of the therapeutic agent produces valuable insight into the leukemogenic mechanism of the genetic lesions. Combinatorial treatment modalities, reflecting

the mutational composition of the leukemic blasts, could be provided in a personalized medicine approach.

In conclusion, genome-wide approaches provided the general view that leukemia is a highly heterogeneous disease conferred by combinations of genetic lesions. NGS enabled the detection of these genetic lesions and demonstrated the existence of clonal dynamics. The major challenges ahead will not relate to the detection of additional genetic lesions but in understanding the functional implications of the acquired genetic abnormalities. Future research will increasingly focus on understanding the leukemogenic mechanism underlying the disease and the dynamics of clonal evolution and at one point should translate this understanding into tailored therapies for AML patients.

REFERENCES

1. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1617-1628.
2. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*. 2009;94(1):131-134.
3. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*. 2004;350(16):1605-1616.
4. Wouters BJ, Lowenberg B, Erpelinck-Verschueren CA, van Putten WL, Valk PJ, Delwel R. 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*. 2009;113(13):3088-3091.
5. Verhaak RG, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood*. 2005;106(12):3747-3754.
6. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006;68(1):49-67.
7. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008;70(1):53-71.
8. Rockova V, Abbas S, Wouters BJ, et al. Risk stratification of intermediate-risk acute myeloid leukemia: integrative analysis of a multitude of gene mutation and gene expression markers. *Blood*. 2011;118(4):1069-1076.
9. Radmacher MD, Marcucci G, Ruppert AS, et al. Independent confirmation of a prognostic gene-expression signature in adult acute myeloid leukemia with a normal karyotype: a Cancer and Leukemia Group B study. *Blood*. 2006;108(5):1677-1683.
10. Eppert K, Takenaka K, Lechman ER, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med*. 2011;17(9):1086-1093.
11. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *JAMA*. 2010;304(24):2706-2715.
12. Metzeler KH, Maharry K, Kohlschmidt J, et al. A stem cell-like gene expression signature associates with inferior outcomes and a distinct microRNA expression profile in adults with primary cytogenetically normal acute myeloid leukemia. *Leukemia*. 2013;27(10):2023-2031.
13. Noordermeer SM, Sanders MA, Gilissen C, et al. High BRE expression predicts favorable outcome in adult acute myeloid leukemia, in particular among MLL-AF9-positive patients. *Blood*. 2011;118(20):5613-5621.
14. Groschel S, Schlenk RF, Engelmann J, et al. Deregulated expression of EVI1 defines a poor prognostic subset of MLL-rearranged acute myeloid leukemias: a study of the German-Austrian Acute Myeloid Leukemia Study Group and the Dutch-Belgian-Swiss HOVON/SAKK Cooperative Group. *J Clin Oncol*. 2013;31(1):95-103.
15. Noordermeer SM, Monteferrario D, Sanders MA, Bullinger L, Jansen JH, van der Reijden BA. Improved classification of MLL-AF9-positive acute myeloid leukemia patients based on BRE and EVI1 expression. *Blood*. 2012;119(18):4335-4337.
16. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med*. 2009;1(6):62.
17. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet*. 2009;10(8):551-564.
18. Sulong S, Moorman AV, Irving JA, et al. A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood*. 2009;113(1):100-107.
19. De Keersmaecker K, Marynen P, Cools J. Genetic insights in the pathogenesis of T-cell acute lymphoblastic leukemia. *Haematologica*. 2005;90(8):1116-1127.

20. Bollag G, Clapp DW, Shih S, et al. Loss of NF1 results in activation of the Ras signaling pathway and leads to aberrant growth in haematopoietic cells. *Nat Genet*. 1996;12(2):144-148.

21. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. *Nature*. 2011;469(7330):343-349.

22. Lee W, Teckie S, Wiesner T, et al. PRC2 is recurrently inactivated through EED or SUZ12 loss in malignant peripheral nerve sheath tumors. *Nat Genet*. 2014.

23. De Raedt T, Beert E, Pasmant E, et al. PRC2 loss amplifies Ras-driven transcription and confers sensitivity to BRD4-based therapies. *Nature*. 2014;514(7521):247-251.

24. Mullighan CG, Goorha S, Radtke I, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007;446(7137):758-764.

25. Mullighan CG, Phillips LA, Su X, et al. Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*. 2008;322(5906):1377-1380.

26. Okamoto R, Ogawa S, Nowak D, et al. Genomic profiling of adult acute lymphoblastic leukemia by single nucleotide polymorphism oligonucleotide microarray and comparison to pediatric acute lymphoblastic leukemia. *Haematologica*. 2010;95(9):1481-1488.

27. Papaemmanuil E, Rapado I, Li Y, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet*. 2014;46(2):116-125.

28. Bassing CH, Swat W, Alt FW. The mechanism and regulation of chromosomal V(D)J recombination. *Cell*. 2002;109 Suppl:S45-55.

29. Shimazaki N, Tsai AG, Lieber MR. H3K4me3 stimulates the V(D)J RAG complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations. *Mol Cell*. 2009;34(5):535-544.

30. Guo C, Yoon HS, Franklin A, et al. CTCF-binding elements mediate control of V(D)J recombination. *Nature*. 2011;477(7365):424-430.

31. Agrawal A, Schatz DG. RAG1 and RAG2 form a stable postcleavage synaptic complex with DNA containing signal ends in V(D)J recombination. *Cell*. 1997;89(1):43-53.

32. Dai Y, Wong B, Yen YM, Oettinger MA, Kwon J, Johnson RC. Determinants of HMGB proteins required to promote RAG1/2-recombination signal sequence complex assembly and catalysis during V(D)J recombination. *Mol Cell Biol*. 2005;25(11):4413-4425.

33. Ji Y, Resch W, Corbett E, Yamane A, Casellas R, Schatz DG. The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. *Cell*. 2010;141(3):419-431.

34. Kharbanda S, Pandey P, Jin S, et al. Functional interaction between DNA-PK and c-Abl in response to DNA damage. *Nature*. 1997;386(6626):732-735.

35. Deutsch E, Dugray A, AbdulKarim B, et al. BCR-ABL down-regulates the DNA repair protein DNA-PKcs. *Blood*. 2001;97(7):2084-2090.

36. Klijn C, Holstege H, de Ridder J, et al. Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res*. 2008;36(2):e13.

37. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61-70.

38. Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346-352.

39. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*. 2013;499(7456):43-49.

40. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.

41. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-421.

42. Egan JB, Shi CX, Tembe W, et al. Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. *Blood*. 2012;120(5):1060-1066.

43. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
44. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219.
45. Narzisi G, O'Rawe JA, Iossifov I, et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat Methods*. 2014;11(10):1033-1036.
46. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-576.
47. Sathirapongsasuti JF, Lee H, Horst BA, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*. 2011;27(19):2648-2654.
48. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *J Genet Genomics*. 2011;38(3):95-109.
49. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
50. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer*. 2011;11(10):726-734.
51. Figueroa ME, Lugthart S, Li Y, et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*. 2010;17(1):13-27.
52. Abdel-Wahab O, Levine RL. Mutations in epigenetic modifiers in the pathogenesis and therapy of acute myeloid leukemia. *Blood*. 2013;121(18):3563-3572.
53. Kon A, Shih LY, Minamino M, et al. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat Genet*. 2013;45(10):1232-1237.
54. Loven J, Hoke HA, Lin CY, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*. 2013;153(2):320-334.
55. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res*. 2011;21(10):1728-1737.
56. Mayba O, Gilbert HN, Liu J, et al. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol*. 2014;15(8):405.
57. Knijnenburg TA, Ramsey SA, Berman BP, et al. Multiscale representation of genomic signals. *Nat Methods*. 2014;11(6):689-694.
58. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med*. 2010;363(25):2424-2433.
59. Mardis ER, Ding L, Dooling DJ, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med*. 2009;361(11):1058-1066.
60. Ribeiro AF, Pratcorona M, Erpelinck-Verschueren C, et al. Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia. *Blood*. 2012;119(24):5824-5831.
61. Roller A, Grossmann V, Bacher U, et al. Landmark analysis of DNMT3A mutations in hematological malignancies. *Leukemia*. 2013;27(7):1573-1578.
62. Chen BF, Chan WY. The de novo DNA methyltransferase DNMT3A in development and cancer. *Epigenetics*. 2014;9(5):669-677.
63. Russler-Germain DA, Spencer DH, Young MA, et al. The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers. *Cancer Cell*. 2014;25(4):442-454.
64. Challen GA, Sun D, Jeong M, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat Genet*. 2012;44(1):23-31.
65. Kim SJ, Zhao H, Hardikar S, Singh AK, Goodell MA, Chen T. A DNMT3A mutation common in AML exhibits dominant-negative effects in murine ES cells. *Blood*. 2013;122(25):4086-4089.

66. Shlush LI, Zandi S, Mitchell A, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*. 2014;506(7488):328-333.

67. Corces-Zimmerman MR, Hong WJ, Weissman IL, Medeiros BC, Majeti R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc Natl Acad Sci U S A*. 2014;111(7):2548-2553.

68. Yasuda T, Ueno T, Fukumura K, et al. Leukemic evolution of donor-derived cells harboring IDH2 and DNMT3A mutations after allogeneic stem cell transplantation. *Leukemia*. 2014;28(2):426-428.

69. Jeong M, Sun D, Luo M, et al. Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet*. 2014;46(1):17-23.

70. Tatton-Brown K, Seal S, Ruark E, et al. Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. *Nat Genet*. 2014;46(4):385-388.

71. Ward PS, Patel J, Wise DR, et al. The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell*. 2010;17(3):225-234.

72. Falini B, Martelli MP, Bolli N, et al. Acute myeloid leukemia with mutated nucleophosmin (NPM1): is it a distinct entity? *Blood*. 2011;117(4):1109-1120.

73. Moran-Crusio K, Reavie L, Shih A, et al. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell*. 2011;20(1):11-24.

74. Figueroa ME, Abdel-Wahab O, Lu C, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell*. 2010;18(6):553-567.

75. Song CX, Szulwach KE, Dai Q, et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*. 2013;153(3):678-691.

76. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011;478(7367):64-69.

77. Greaves M, Maley CC. Clonal evolution in cancer. *Nature*. 2012;481(7381):306-313.

78. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*. 2012;481(7382):506-510.

79. Walter MJ, Shen D, Ding L, et al. Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med*. 2012;366(12):1090-1098.

80. Wong TN, Ramsingh G, Young AL, et al. Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature*. 2014.

81. Chan SM, Majeti R. Role of DNMT3A, TET2, and IDH1/2 mutations in pre-leukemic stem cells in acute myeloid leukemia. *Int J Hematol*. 2013;98(6):648-657.

82. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014;11(4):396-398.

83. Ha G, Roth A, Khattra J, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res*. 2014.

84. Wang Y, Waters J, Leung ML, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*. 2014;512(7513):155-160.

85. Fluidigm. http://www.fluidigm.com/c1wes.html.

86. Shalek AK, Satija R, Shuga J, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*. 2014;510(7505):363-369.

87. Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res*. 2014;42(14):8845-8860.

88. Smallwood SA, Lee HJ, Angermueller C, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods*. 2014;11(8):817-820.

89. Urnov FD, Miller JC, Lee YL, et al. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*. 2005;435(7042):646-651.

90. Joung JK, Sander JD. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol*. 2013;14(1):49-55.

91. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012;337(6096):816-821.

92. Soldner F, Laganiere J, Cheng AW, et al. Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell*. 2011;146(2):318-331.

93. Maddalo D, Manchado E, Concepcion CP, et al. In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature*. 2014.

94. Genovese P, Schiroli G, Escobar G, et al. Targeted genome editing in human repopulating haematopoietic stem cells. *Nature*. 2014;510(7504):235-240.

95. Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*. 2014.

96. Sanchez-Rivera FJ, Papagiannakopoulos T, Romero R, et al. Rapid modelling of cooperating genetic events in cancer through somatic genome editing. *Nature*. 2014.

97. Shi J, Vakoc CR. The mechanisms behind the therapeutic activity of BET bromodomain inhibition. *Mol Cell*. 2014;54(5):728-736.

98. Baines AT, Xu D, Der CJ. Inhibition of Ras for cancer treatment: the search continues. *Future Med Chem*. 2011;3(14):1787-1808.

99. Downward J. Targeting RAS signalling pathways in cancer therapy. *Nat Rev Cancer*. 2003;3(1):11-22.

100. Fathi AT, Chen YB. Treatment of FLT3-ITD acute myeloid leukemia. *Am J Blood Res*. 2011;1(2):175-189.

101. Fenaux P, Mufti GJ, Hellstrom-Lindberg E, et al. Azacitidine prolongs overall survival and reduces infections and hospitalizations in patients with WHO-defined acute myeloid leukaemia compared with conventional care regimens: an update. *Ecancermedicalscience*. 2008;2:121.

102. Gurion R, Vidal L, Gafter-Gvili A, et al. 5-azacitidine prolongs overall survival in patients with myelodysplastic syndrome--a systematic review and meta-analysis. *Haematologica*. 2010;95(2):303-310.

103. Wang F, Travins J, DeLaBarre B, et al. Targeted inhibition of mutant IDH2 in leukemia cells induces cellular differentiation. *Science*. 2013;340(6132):622-626.

104. Rohle D, Popovici-Muller J, Palaskas N, et al. An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. *Science*. 2013;340(6132):626-630.

105. Daigle SR, Olhava EJ, Therkelsen CA, et al. Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. *Cancer Cell*. 2011;20(1):53-65.

APPENDIX

Nederlandse samenvatting

Hematopoëse is een continue proces waarbij dagelijks verschillende bloedcellen in zeer grote aantallen aangemaakt worden. Er zijn een aantal verschillende typen bloedcellen en elk type heeft een specifieke functie in het lichaam. Witte bloedcellen, onderverdeeld in granulocyten, monocyten, macrofagen en lymfocyten, spelen een belangrijke rol in de verdediging van het lichaam tegen pathogene indringers, terwijl rode bloedcellen een belangrijke rol spelen bij het vervoeren van zuurstof door het lichaam en bloedplaatjes bij bloedstolling. Al deze bloedcellen ontstaan uit één type niet-gespecialiseerde cellen, genaamd "hematopoietische stamcellen", in het beenmerg. Deze hematopoietische stamcellen hebben de capaciteit om voldoende functionele bloedcellen te produceren naar gelang de fysiologische behoeften van het lichaam. Hematopoietische stamcellen produceren dochtercellen welk verder uitrijpen door een proces genaamd "celdifferentiatie". Na uitrijping worden deze bloedcellen vrijgelaten in het perifere bloed vanwaar zij hun functionele eigenschappen kunnen uitvoeren.

Leukemie, ook wel bloedkanker genoemd, is een ziekte van het beenmerg. De uitrijping van één type bloedcel is verstoord en de ongelimiteerde vermenigvuldigingen van deze cel leidt tot de complete overwoekering van het beenmerg en leukemische cellen in het bloed. De opeenhoping van deze niet-functionele leukemische cellen in het beenmerg verstoort de normale productie van bloedcellen wat leidt tot de bijna complete afwezigheid van functionele bloedcellen in het bloed. In dit proefschrift is er onderzoek gedaan naar acute myeloïde leukemie (AML) en in mindere mate acute lymfatische leukemie (ALL). Acute leukemie wordt onder andere gekarakteriseerd door genetische afwijkingen in het DNA. Door klinisch wetenschappelijk onderzoek is aangetoond dat bepaalde genetische afwijkingen associëren met een goede prognose, terwijl andere minder goede vooruitzichten bieden. De afgelopen jaren zijn er grote vorderingen geboekt in de kennis aangaande verworven mutaties en hun (klinische) associatie met acute leukemie. Uit deze onderzoeken is gebleken dat AML, maar ook ALL, een verzameling van aandoeningen is, waarbij verschillende genetische afwijkingen ten grondslag liggen. Deze acute leukemie subtypen worden geïdentificeerd aan de hand van verworven genetische afwijkingen in combinatie met klinische en cytomorfologische eigenschappen. De correcte bepaling van acute leukemie subtypen is van uitermate belang voor de juiste behandeling en prognosestelling en behoeft verder bestudering.

Recente technologische ontwikkelingen hebben ervoor gezorgd dat het genetische materiaal van AML patiënten in detail gekarakteriseerd kunnen worden en hebben geleid tot een verbeterde bepaling van de verschillende AML subtypen. De vooruitgang in technologische methodieken stelt de onderzoeker in staat om veel verschillende genetische eigenschappen van de leukemische cellen te bepalen, denkend aan miljoenen metingen, met bijkomend nadeel dat de data onoverzichtelijk wordt voor standaard empirische waarnemingen. De oplossing tot dit probleem behoeft het gebruik van computationele biologie of statistische modelering om de meeste waardevolle data uit te lichten. Het werk beschreven in dit proefschrift richt zich op de identificatie van genetische en epigenetische afwijkingen in AML aan de hand van genoombrede

technologieën. De nadruk ligt vooral op het gebruik van drie verschillende genoom-brede technologieën: (1) genexpressie microarrays voor gen expressie bepalingen, (2) DNA mapping arrays voor het detecteren van genetische afwijkingen; vooral kopieaantal verschillen, (3) next generation sequencing (NGS) voor het vaststellen van genetische en epigenetische afwijkingen. In dit proefschrift werden deze genoom-brede technologieën, soms in combinatie, gebruikt voor de identificatie van nieuwe AML subtypen of om meer inzicht te krijgen in het mechanistische proces ten grondslag aan de ontwikkeling van leukemie.

In **hoofdstuk 2** hebben wij een statistische methode ontwikkeld die gebruikt maakt van het multinomiale logistische regressie model in combinatie van een aangepaste groep lasso regularisatie procedure. Deze methode produceert ijle genexpressie predictie patronen waarmee de participatiekans per patiënt voor verschillende klassen tegelijkertijd bepaald kan worden. Genexpressie predictie patronen, bestaand uit een gelimiteerd aantal genen, worden voor alle vooraf bepaalde klassen bepaald. In elk genexpressie predictie patroon wordt een gen gewogen aan de hand van zijn invloed op de klasse-predictie. Wij hebben vastgesteld, vergelijkbaar aan voorgaande studies, dat AML subtypen met klinisch gunstige cytogenetische afwijkingen, zoals t(15;17), t(8;21) en inv(16), accuraat ingedeeld kunnen worden. Verder hebben wij aangetoond dat onze statistische methode een verbeterde indeling geeft voor AML patiënten met een combinatie van mutaties in de genen *NPM1* en *FLT3*. In **hoofdstuk 3** hebben wij onderzocht wat de klinische relevantie is van *CEBPA* enkel- of dubbelmutanten, genaamd *CEBPA*^{sm} en *CEBPA*^{dm}. Daarbij is onderzocht of deze mutaties geassocieerd zijn met een specifieke genexpressie patroon of gekarakteriseerd worden door een specifiek mutatie spectrum. *CEBPA*^{dm} AML patiënten hebben weinig extra recurrente mutaties, terwijl *CEBPA*^{sm} AML patiënten veelal mutaties hebben in de genen *NPM1* en *FLT3*. Vervolgens stelden wij vast dat de genetische marker *CEBPA*^{sm} geen prognostische waarde heeft, terwijl de genetische marker *CEBPA*^{dm} een sterke associatie heeft met een gunstig klinisch vooruitzicht. Gebruikmakend van een multinomiale logistische regressie model met een lasso regularisatie procedure, vergelijkbaar aan het model beschreven in **hoofdstuk 2**, hebben wij aangetoond dat *CEBPA*^{dm} AML patiënten accuraat voorspeld kunnen worden in een AML validatiecohort, terwijl dit niet mogelijk is voor *CEBPA*^{sm} patiënten.

In **hoofdstukken 4 en 5** hebben wij onderzocht welke genetische afwijkingen aanwezig zijn in AML en ALL patiënten door het gebruik van DNA mapping arrays. In **hoofdstuk 4** beschrijven wij de ontwikkeling van een softwarepakket voor het visualiseren van data afkomstig van DNA mapping arrays in combinatie met genexpressie data. DNA kopieaantal veranderingen, op dezelfde plaats in het genoom, kunnen worden gebruikt als gids voor het identificeren van genen met een belangrijke rol bij de ontwikkeling van leukemie. Voor het bepalen van het verlies van heterogeniteit voor AML patiënten zonder beschikbaar genetisch materiaal van remissie of normaal weefsel hebben wij een nieuw statistisch model ontwikkeld. In **hoofdstuk 5** onderzochten wij terugkerende genetische afwijkingen in AML en ALL patiënten doormiddel van DNA mapping arrays en NGS technologieën. Genetisch materiaal van ALL patiënten vertonen

relatief veel kleine verliezen (deleties) of amplificaties in het DNA, terwijl genetisch materiaal van AML patiënten weinig (terugkerende) genetische afwijkingen vertonen. Sommige genetische afwijkingen zijn ALL-breed, terwijl andere genetische afwijkingen heel specifiek zijn voor B-ALL of T-ALL. Opvallend hierbij was dat alle T-ALL patiënten deleties van de genen *CDKN2A/B* of genen geassocieerd met *CDKN2A/B* hadden verworven. Daarnaast ontdekten wij dat de proximale genen *NF1* en *SUZ12* samen afwijkend zijn in T-ALL en AML subgroepen. De aangetoonde afwijking veroorzaakt een significante verlaging van de genexpressie levels van beide genen. Opvallend is dat er vaak mutaties gevonden worden op het overgebleven normale *NF1* allel. In samenspraak met voorgaande studies postuleren wij dat het verlies van NF1, een remmer van het oncogen *RAS*, samenwerkt met het verlies van het polycomb repressive complex 2. Een interessante bevinding is dat specifieke B-ALL patiënten een tendens vertonen voor het overmatig genetisch verlies van gen promotors of een gedeelte van een gen. DNA sequentiebepalingen van regio's die de deleties flankeren in het genetisch materiaal van 5 B-ALL patiënten, door middel van NGS, demonstreerde de aanwezigheid van zogeheten cryptische recombinatie signaalsequenties (RSSs). Deze sequentiemotieven waren aanwezig aan één of beide zijdes voor 91% van de verworven deleties en bovendien werden er vaak willekeurige nucleotiden aan de breekpunten toegevoegd. Sequentiemotief zonder enige vorm van voorkennis toonde aan dat meeste deletiebreekpunten worden gekarakteriseerd door 12-basepaar-afstand RSS en 23-basepaar-afstand RSS sequentie motieven. Dit impliceert dat de deleties het gevolg zijn van onrechtmatige genetische herschikkingen door het recombinatie-activerende genen (RAG) eiwitcomplex. Epigenetische data analyses op data afkomstig van een B-lymfocytaire cellijn toonde aan dat de deletiebreekpunten verrijkt zijn voor de epigenetische veranderingen H3K4me3 en H3K27ac, en de binding van RNA polymerase II. Deze epigenetische veranderingen zijn vergelijkbaar aanwezig in gebieden welk immuunreceptor-herschikkingen ondergaan door het RAG eiwitcomplex. De vertaling van de humane breekpuntlocaties naar homologe gebieden in het muisgenoom toonde aan dat deze locaties verrijkt zijn voor Rag2 binding gedetecteerd in muis thymocyten. Naast de inductie van deleties kan het RAG eiwitcomplex zogeheten open-en-dicht aberraties veroorzaken door gebruik te maken van foutgevoelige DNA-herstelmechanismen. Dit mechanisme is momenteel het onderwerp van actieve bestudering

De **hoofdstukken 6 tot 10** richten zich op het onderzoek aan genetische en epigenetische veranderingen, terwijl **hoofdstuk 11** zich richt op de ontwikkeling van een nieuwe statistische methodologie voor het schatten van kopieaantal veranderingen aan de hand van NGS data. In **hoofdstuk 6**, worden de genetische afwijkingen in de sequentiële bloed- of beenmergmonsters van een ernstig aangeboren neutropenie (SCN) patiënt bepaald die uiteindelijk na 17 jaar G-CSF behandeling leukemie ontwikkelt. In totaal werden er 12 verworven mutaties gevonden in de leukemische fase, waarvan 3 van deze mutaties al aanwezig waren in een laag aantal cellen jaren voordat de leukemie ontwikkelt. Eén van deze mutaties was verworven in het gen dat de G-CSF receptor codeert, daarnaast werd er in de leukemische fase nog een G-CSF receptor

mutatie ontdekt. Deze laatst verworven G-CSF receptor mutatie maakt dat de leukemische cellen onafhankelijk worden van het G-CSF cytokine. In **hoofdstuk 7** onderzochten wij hoe afwijkingen op de lange arm van chromosoom 3, zoals inv(3)(q21q26.2) of t(3;3)(q21;q26) (afgekort inv(3)/t(3;3)), een overexpressie van het proto-oncogen *EVI1* veroorzaken. Inv(3)/t(3;3) DNA breekpuntbepalingen toonde aan dat een specifiek gebied altijd wordt gerepositioneerd naar het chromosomale 3q26 gebied. De integratie van RNA-Seq, Chip-Seq en 4C-Seq data demonstreerde dat er een *GATA2* enhancer aanwezig is in het gerepositioneerd gebied. Deze gerepositioneerde enhancer ondergaat een interactie met de *EVI1* promotor met overexpressie als gevolg. Het *GATA2* gen verliest een essentiële enhancer op het afwijkende chromosoom, waardoor *GATA2* alleen nog maar tot expressie komt vanaf het resterende normale chromosoom. Inactivatie van de gerepositioneerde enhancer in de cellijn MUTZ-3 veroorzaakt compleet verlies van *EVI1* expressie. Tijdens het repositioneren van het chromosomale 3q21 gebied vormt er een super-enhancer gekarakteriseerd door de verrijking van de epigenetische verandering H3K27ac. Behandeling van de cellijn MOLM1 en MUTZ-3 met een BET-bromodomein remmer (JQ1) veroorzaakt *EVI1* genexpressie verlies, vergelijkbaar met eerdere experimenten. In **hoofdstuk 8** hebben wij verworven mutaties in kaart gebracht voor myeloïde maligniteiten met inv(3)/t(3;3) cytogenetische afwijkingen. In 98% van de inv(3)/t(3;3) patiënten worden mutaties in genen van het RAS/RTK signaaltransductie-netwerk verworven. Heterozygote mutaties werden gevonden in het resterende normale *GATA2* allel en de genen *SF3B1* en *RUNX1*. Verschillende verworven mutaties werden gevonden in genen coderend voor epigenetische modificeerders. Een interessante bevinding is dat mutatie noch gen expressie patronen kunnen differentiëren tussen verschillende inv(3)/t(3;3) myeloïde maligniteiten, suggererend dat deze maligniteiten mogelijk als één ziekte beschouwd kunnen worden. In **hoofdstuk 9** detecteren wij een nog onbekend *KMT2A-MYH11* fusietranscript in een patiënt die na een myelodysplastische fase uiteindelijk AML ontwikkelt. Interessant is het feit dat dit fusietranscript al aanwezig was in de myelodysplastische fase. In **hoofdstuk 10** beschrijven wij AML patiënten met een verworven extra kopie van het gen *BCL11B*. De extra kopie is het gevolg van een springende translocatie waarbij het *BCL11B* locus integreert in super-enhancers gelegen op andere chromosomen, met *BCL11B* overexpressie tot gevolg. In **chapter 11** beschrijven wij de ontwikkeling een nieuw statistische methode voor het schatten van DNA kopieaantal veranderingen aan de hand van NGS data. Deze methode bepaalt de aanwezigheid van systematisch statistische ruis in een referentie dataset van gezonde diploïde (twee kopieën per chromosoom) individuen. De bepaalde systematisch statistische ruis wordt vervolgens gebruikt voor het accuraat schatten van DNA kopieaantal veranderingen in chromosomen vanuit NGS verkregen data. In vergelijking met de traditioneel gebruikte DNA mapping arrays kunnen wij met onze statistische methode meer en preciezer gevalideerde DNA kopieaantal veranderingen detecteren. In het afsluitende **hoofdstuk 12** worden de meest belangrijke bevindingen en hypothesen uit dit proefschrift in een breder context behandeld.

APPENDIX

Dankwoord

"Of all the things which wisdom provides to make us entirely happy,
much the greatest is the possession of friendship."

Epicurus

Beste lezers, jaren van onderzoek zijn aan dit proefschrift voorafgegaan. Met veel plezier en trots kan ik zeggen dat mijn boekje voltooid is. Het voltooien van mijn proefschrift zou zonder de hulp van vele mensen nooit gelukt zijn. Daarvoor maak ik gebruik van dit laatste, edoch meest gelezen gedeelte van het proefschrift, om een aantal mensen persoonlijk te bedanken.

Mijn eerste woorden richt ik tot mijn promotor, **Bob Löwenberg**. Als beginnend bachelor student mocht ik op jouw afdeling te werk gaan aan verschillende opeenvolgende projecten en uiteindelijk als PhD-student onder jouw hoede. Je stond altijd open voor mijn ideeën en wij hebben vele projecten besproken in jouw kamer achterin de kopkamer. Jouw kritische, maar opbouwende, kanttekeningen gemaakt tijdens mijn presentaties zijn altijd ten faveure geweest voor de kwaliteit van mijn werk. Beste **Bob**, bedankt voor jouw betrokkenheid, visie en steun. Het was mij een waar genoegen om jou als promotor gehad te hebben.

Tijdens mijn PhD-traject is een additionele promotor toegevoegd, **Ruud Delwel**. Beste **Ruud**, ik wil jou bij deze bedanken voor jouw persoonlijke en intensieve begeleiding, zeker gedurende de laatste maanden tijdens de afronding van dit proefschrift. Op bijna elk moment van de dag was jij beschikbaar, waardoor ik altijd kon binnenstormen als ik iets nieuws had te vertellen. Met een grap wist jij mijn soms wat iets te naïeve en serieuze houding te doorbreken. Zeker als mijn mentor gedurende de volgende stap van mijn onderzoekscarrière zullen wij nog aan vele projecten samenwerken.

Specifieke aandacht wil ik schenken aan mijn co-promotor, **Peter Valk**. Beste **Peter**, ik had mij geen betere begeleider kunnen toewensen. Als bachelor student ben ik in jouw groep begonnen en 10 jaar later eindig ik mijn PhD-traject onder jouw hoede. Het bellen tijdens jouw vakanties heb ik toch enigszins gemodereerd. We hebben, wegens congressen, veel reizen naar het buitenland gemaakt en tijdens deze reizen hebben wij een voorkeur ontwikkeld om elke hoge toren te beklimmen met een niet geringe kans dat ik weer eens van de trap val tijdens de afdaling. Ik heb zeer genoten van het Peter-Ruud één-tweetje, waarbij de grappen steeds flauwer worden gecorreleerd aan de hoeveelheid geconsumeerd bier. Vanuit mijn PhD-traject hebben wij nog projecten ten overvloede om de komende jaren met plezier te vullen. Bedankt voor jouw intensieve steun, betrokkenheid, input, drive en de mogelijkheid om op elk moment van de dag bij jou langs te komen.

Beste **prof. dr. Jelle Goeman, prof. dr. Sjaak Philipsen** en **prof. dr. Gert Ossenkoppele** bedankt voor jullie deelname in de beoordelingscommissie van mijn proefschrift. Beste **Jelle**, bedankt voor jouw begeleiding, inzet en motivatie. Je bent een toonbeeld voor statistici met een verfrissende nieuwsgierigheid in de biologische kant van het verhaal. Jouw begeleiding heeft mij geholpen bij mijn statistische kennisontwikkeling. Onze reizen naar Frankrijk voor verschillende statistische congressen zal ik nooit vergeten. Op zoek naar een restaurant waar geen, maar dan ook zeker geen, toeristen komen in de zijstraten van Parijs. De Lamprei-schotel in Bordeaux zag er toch wat minder appetijtelijk uit. Ik wens je heel veel succes als nieuwbakken professor statistiek in Nijmegen en ik hoop nog op vele projecten samen te werken. Beste **Sjaak** en **Gert**, bedankt voor jullie deelname en inzet in de kleine commissie.

Beste **prof. dr. Ivo Touw**, **prof. dr. Wouter den Laat** en **dr. Bert van der Reijden**, bedankt voor jullie deelname aan de verdediging van mijn proefschrift. Beste **Ivo**, ook wij werken alweer 10 jaar samen. Ik weet nog goed dat ik bij jou op de kamer zat voor een interview over hoe wij MADEx konden verbeteren. Jouw aanwezigheid en kritische, doch opbouwende, opmerkingen zijn altijd zeer gewaardeerd en waardevol geweest. Bedankt voor jouw inzet, advies en motivatie. Beste **Wouter**, bedankt voor de deelname. Zonder jouw 4C-Seq technologie was onze Cell paper nooit tot stand gekomen. Bedankt dat wij de 4C-Seq technologie van jouw groep mochten leren. Beste **Bert**, het was mij een waar genoegen om op een aantal projecten samen te werken. Ik heb altijd zeer genoten van jouw advies en inbreng.

De (ex-)leden van de Valk/Rijneveld research groep; **François**, **Annelieke**, **Adil**, **Jasper**, **Anikó** en **Carla**. **François**, man van het zuiden. Ik heb met heel veel plezier met jou samengewerkt. Jouw rigoureuze en accurate handelswijze is bewonderenswaardig. Mocht Dries Roelvink in de buurt zijn dan gaan we er samen heen. **Annelieke**, bedankt voor de vele experimenten die jij voor mij in het verleden hebt uitgevoerd en alvast bedankt voor de vele die nog gaan volgen. Ik vind het noemenswaardig dat je niet terugdeinst voor het uitvoeren van volledig onbekende experimenten (ook binnen de afdeling). A brief switch to English to address our friend from Oman, **Adil**. Adil, the latest addition to the Valk group, your journey has just started and I wish you the best. I already observed great potential in your work and scientific career ahead. Soon-to-be father, I wish you all the best. You'll notice it will be tough in the beginning but also the best time of your life. **Jasper**, metal-gast en Euphorbia Leuconeura-man. Een tijdje geleden de Valk groep ingewisseld voor de Sonneveld groep, maar niet vergeten. Wij hebben een gedeelde passie voor metal muziek en tuinieren, toch wel een beetje vreemde combinatie als je het mij vraagt. Zodra Insomnium weer in het land is zullen wij elkaar daar, in de concertzaal, weer ontmoeten. **Anikó**, recent begonnen in de Rijneveld groep en nog iemand die mijn passie voor tuinieren deelt. Mijn deur staat altijd open voor NGS-vragen! **Carla**, bedankt voor alle hulp die je mij geboden hebt! Heel veel plezier met de kleine!

De dames van de Valk diagnostiek groep; **Isabel**, **Wendy**, **Pauline**, **Sonja**, **Marloes, Chantal** en **Antoinette**. Ik wil jullie bedanken voor jullie ondersteuning tijdens het voltooien van mijn onderzoek en proefschrift. **Isabel**, rooibos thee is niet te drinken! **Wendy**, de diagnostiek backup staat veilig! **Pauline**, veel succes met toekomstige wielrenklassiekers! **Sonja**, ik zal de klinische tool binnenkort repareren! **Marloes**, heel veel succes in Utrecht binnenkort! **Chantal**, rij niet te hard op de motor! **Antoinette**, succes in Groningen en ik wens je heel veel plezier op de boerderij!

Mijn (ex-)mede-bioinformatici: **Remco**, **Rowan**, **Erdogan** en niet te vergeten **Roel Verhaak**. Beste **Remco**, wij kennen elkaar al sinds de TU Delft waar wij een overgroot gedeelte van het curriculum samen hebben gevolgd. Ik vind het nog steeds geweldig dat je bij ons bent komen werken, ondanks je soms wat rare grappen en grollen (*kuch* plakband over de sensor van mijn muis *kuch*). Ik hoop nog vele jaren samen te werken. **Rowan**, de jongste van het stel. Het was super dat je het laatste jaar ons kwam vergezellen op de kamer. Ik wens je nog heel veel succes toe met het afronden van jouw proefschrift en misschien krijgen wij dan eens de veel beloofde appeltaart. **Erdogan**, de Circos plot-sensei. Ik heb genoten van onze samenwerking en discussies over totaal irrelevante onderwerpen. Heel veel succes met de voortzetting van jouw werk bij de TU Delft. **Roel**, mijn voorganger en voorgaande mentor. Dit proefschrift is een voortzetting van jouw en Peter's werk. De jaren voor mijn PhD-traject zijn een goede voorzet geweest tot waar ik nu ben. Recent nog gesproken en ik weet dat je nog steeds van metal houdt. Zelfs nu je ouder en wijzer bent en kinderen hebt.

Leden van de Delwel groep; **Stefan G.**, **Roberto**, **Marije**, **Claudia** en **Eric**. **Stefan G.**, the man with the golden hands. Without your excellent scientific and lab skills the *EVI1* project wouldn't have gotten as far as it does today. I'm still impressed at what you achieved during just a few years in Rotterdam. I sincerely apologies for the frequent midnight calls during the Cell paper submission and revision. I still have a bottle of Château Haut-Bages Libéral at home that we need to finish. I bid you good luck in Heidelberg and I'm confident you'll achieve great scientific wonders. **Roberto**, you eat, sleep and breathe *CEBPA*. I'm confident that "peak 6", also dubbed the black hole-enhancer, will provide you with sufficient scientific material for years to come. I thank you for all the discussions we had on a wide variety of scientific and non-scientific topics! **Claudia** en **Marije**, heel erg bedankt voor de ondersteuning die jullie geleverd hebben aan onze projecten. Zonder jullie was het nooit gelukt. **Eric**, de *EVI1*-man. De "go-to man" voor al uw vragen over *EVI1*, lab experimenten of next generation sequencing. Heel erg bedankt voor alle inzet en adviezen tijdens mijn projecten. Jouw inbreng tijdens mijn werkpresentaties wordt altijd zeer gewaardeerd.

De P.I.'s; **Anita**, **Mojca**, **Marc**, **Stefan E.**, **Jan C.**, **Eric B.**, **Frank**, **Moniek**, **Dick, Tom** en **Pieter**. Ik wil jullie bij deze bedanken voor jullie contributie aan dit proefschrift en niet te vergeten de vele opmerkingen en adviezen verkregen tijdens de werkbesprekingen. Ik wens jullie en de projecten onder jullie hoede het beste toe.

Dear **Elwin**, **Su Ming**, **Bas**, **Sanne**, **Renée**, **Jurgen**, **Eric V.**, **Ferry**, **Julien**, **Piotr**, **Adrian**, **Saman**, **Mark van Duin**, **Paulette**, **Hans de Looper**, **Annemiek**, **Tomasia**, **Menno**, **Anita S.**, **Peter van Geel**, **Niken**, **Michael**, **Natalie**, **Martijn** and **Egied**, thank you for all your help! Dear **Noemi**, **Kasia, Si**, **Zhen**, **Keane**, **Avinash**, **Davine**, **Julia**, **Patricia O.**, **Patricia D.**, **Monica**, **Michelle**, **Farshid** and **Roel P.,** I wish you the best of luck with writing your theses!

De dames van het secretariaat; **Leenke, Annelies** en **Ans**. **Leenke**, bedankt voor alle hulp tijdens mijn PhD-traject. Ik ben me ervan bewust dat ik soms erg verstrooid kan zijn. **Annelies**, bedankt voor alle hulp gedurende het einde van mijn PhD-traject. Zonder jou was deze dag nooit tot stand gekomen. **Ans**, recent met pensioen gegaan, wil ik alsnog bedanken. Ik hoop dat je ondertussen al een leuke hobby hebt kunnen vinden om je vrije tijd aan te besteden.

Beste **Jan van Kapel**, man van de computers, software en het jagen. Ik wil je bedanken voor de technische ondersteuning die je de afgelopen jaren hebt geboden.

Beste **(ex-)collegae van de afdeling medische statistiek en bio-informatica in Leiden**, ik wil jullie allemaal bedanken voor jullie input, voor de leuke discussies op wetenschappelijk en filosofisch gebied en de tijd dat ik bij jullie heb mogen spenderen. Beste **Rosa**, bedankt voor alle hulp en de leuke discussies die wij gevoerd hebben. Veel succes met het afronden van het PhD-traject.

Most scientific progress results from fruitful collaborations and I wouldn't do justice without thanking all scientific collaborators. **Lars Bullinger**, **Konstanze Döhner** and **Hartmut Döhner (**Ulm university), I'm indebted to you for all the valuable input given during the last years and AML samples sent in the different projects described in this thesis. I had a great time visiting your department this year and hope that we could collaborate on different projects in the future. **Berna Beverloo** (afdeling Klinische Genetica), bedankt voor alle cytogenetische bepalingen en het verzorgen van correcte cytogenetische annotatie. **Kirsten van Lom** (afdeling Hematologie), bedankt voor alle cytomorfologische bepalingen, thee en advies. **Vincent van der Velden** en **Ton Langerak** (afdeling Immunologie), bedankt voor alle bepalingen. **Harmen van der Werken** (afdeling Urologie), dank voor jouw hulp met het verwerken van 4C-Seq data. **Elzo de Wit** en **Britta Bouwman** (Hubrecht Institute), dank voor jullie hulp voor het opzetten van de 4C-Seq technologie binnen onze afdeling. **Joop Jansen** (afdeling Hematologie, Raboud University Medical Centre), dank voor jouw advies en inzet. **Marta Pratcorona** (Hospital Clínic de Barcelona) helped

in setting up the research described in Chapter 5. I wish you all the best back in Barcelona! Dank voor alle deelnemende HOVON-centra – zonder hen was dit proefschrift niet tot stand gekomen. **Wim van Putten** en **Yvette van Noorden**, bedankt voor de up-to-date patiënteninformatie en wetenschappelijke input. **Jonas Jutzi** and **Heike Pahl** (Freiburg University), I really enjoyed your scientific input and time spent together in Rotterdam and Ulm. Hopefully, I could visit Freiburg once and enjoy your hospitality.

Beste **(oud-)collegae van de afdeling hematologie**, ik wil jullie bedanken voor jullie tomeloze inzet en wetenschappelijke input, voor alle gezellige tijden samen gespendeerd, leuke discussies en alle hulp die ik van jullie heb mogen ontvangen.

Mijn paranimfen **Jeffrey van Heck** en **Mark van den Berg**. Het dynamische trio is weer eens samengekomen. Al vrienden sinds de kleuterschool en nog steeds goed bevriend na zoveel jaren. Ik ben dankbaar voor jullie hulp om deze dag een succes maken. **Jeffrey**, alweer wat jaartjes verhuisd naar Hilversum, maar nog steeds in goed contact. Je bent een goede vriend en een echte levensgenieter. Vroeger leken wij heel erg op elkaar, waardoor mensen ons nooit uit elkaar konden houden. Recent getrouwd, wat toch als een verrassing kwam, omdat je het nooit wilde doen. Ik waardeer je gesprekken, inzet en levensinstelling enorm. Ik hoop nog vele avonturen samen te beleven. **Mark**, tevens hechte vriend en tennis-mattie. We kennen elkaar al sinds de kleuterschool waar we vaak bij elkaar thuis verbleven. Bovendien hebben wij jaren samen tennis gespeeld, iets wat je nog steeds fanatiek doet. Ik vind het geweldig dat je mij deze dag bij wilt staan en moge wij nog vele jaren bevriend blijven! Lieve **vrienden**, dank jullie voor alle bijeenkomsten, feestjes, vrijgezellenfeestjes, de vele avonden samen en de goede gesprekken.

Lieve **familie**, ik wil jullie bedanken voor al jullie steun gedurende de afgelopen jaren. **Ton**, **Hannie**, **Tom**, **Lisa**, **opa en oma Smit**, bedankt voor jullie ondersteuning gedurende het verhuizen, de geboorte en de drukke dagen. Ik weet dat Jonathan altijd graag bij jullie op bezoek komt.

Lieve grote zus **Jorunn**, als kleine broer moet ik toch altijd naar je opkijken. Je hebt je eigen weg gevolgd, bent meester in de rechten geworden en daarmee altijd gelukkig geweest. Afgelopen jaren de moeder geworden van **Sana**, **Isra** en **Maysa**, drie prachtige dochters. **Nabil**, je bent een geweldige zwager. Misschien moet je mij binnenkort toch maar weer eens leren autorijden.

Lieve **pa** en **ma**, ik heb enorme bewondering voor jullie. Altijd druk bezig geweest met jullie bedrijf en de enige twee mensen die ik ken die zoveel gewerkt hebben. Van jongs af aan hebben jullie mij en Jorunn altijd een vrije keuze gegeven en wij hebben ons nooit hoeven te bekommeren of de mogelijkheid tot deze keuze wel bestond. Geluk moesten wij zelf maken en ons hart volgen en dat hebben wij ook gedaan. Ik dank jullie voor de steun in alle jaren, zonder jullie was dit proefschrift nooit tot stand gekomen.

Mijn laatste woorden wil ik uiteraard wenden aan mijn lieve **Kristel** en **Jonathan**. Lieve **Kristel**, wij zijn alweer tien jaar ontzettend gelukkig samen en zijn daarom recent getrouwd. Jij maakt mij compleet en bent een rots in de branding van mijn toch iets chaotische leven. Daarnaast ben je een lieve en zorgzame moeder die alles voor onze lieve zoon over heeft. Bedankt dat je er altijd voor mij bent! Lieve **Jonathan**, helaas kan je nog niet lezen, maar ooit ergens in de toekomst zal je dit "dan oude" proefschrift oppakken en het misschien wel begrijpen. Naast de ontmoeting van jouw moeder ben jij het mooiste wat mij is overkomen! Wie weet, in de toekomst, kunnen wij dit proefschrift met elkaar bediscussiëren.

Mathijs

APPENDIX

Curriculum vitae

De auteur van dit proefschrift werd op 24 augustus 1984 geboren in Rotterdam. Na het afronden van het Hoger Algemeen Voortgezet Onderwijs aan het Gemini College in Ridderkerk startte hij de studie Hogere Informatica aan de Hogeschool Rotterdam in september 2002. Hij voltooide vier jaar later de opleiding met een afstudeeronderzoek getiteld, "*Cluster analyse van genoom-brede gene expressie profielen en het visualiseren van genoom-brede SNP patronen*", onder supervisie van dr. Roel G.W. Verhaak en dr. Peter J.M. Valk op de afdeling hematologie van het Erasmus Universitair Medisch Centrum (Erasmus MC). Vervolgens begon hij in september 2006 aan de opleiding Bio-informatica aan de Technische Universiteit te Delft (TU Delft). Als masterexamen (doctoraalexamen) voerde hij aan het Leids Universitair Medisch Centrum (LUMC), onder supervisie van dr. Jelle J. Goeman en prof. dr. Marcel J.T. Reinders, 9 maanden onderzoek naar specifieke gen expressie predictie patronen verkregen door de integratie van een gemodificeerde groep lasso procedure in multinomiale logistische regressie modellen. Na het *cum laude* behalen van het masterexamen begon hij in december 2009 als promovendus in de onderzoeksgroep van dr. Peter J.M. Valk op de afdeling hematologie van het Erasmus MC (promotoren prof. dr. Bob Löwenberg en prof. dr. Ruud Delwel). Aldaar vond het onderzoek beschreven in dit proefschrift plaats. In december 2014 begon hij als postdoctoraal onderzoeker op de afdeling hematologie van het Erasmus MC (prof. dr. Ruud Delwel).

APPENDIX

Publications

1 **Sanders MA**, Kavelaars FG, Zeilemaker A, Al Hinai AS, Abbas S, Beverloo HB, van Lom K, and Valk PJ, RNA sequencing reveals a unique fusion of the lysine (K)-specific methyltransferase 2A and smooth muscle myosin heavy chain 11 in myelodysplastic syndrome and acute myeloid leukemia, Haematologica*,* 100 (2015), e1-e3l.

2 Gröschel S*, **Sanders MA***, Hoogenboezem R, Zeilemaker A, Havermans M, Erpelinck C, Bindels EM, Beverloo HB, Dohner H, Lowenberg B, Dohner K, Delwel R, and Valk PJM, Mutational spectrum of myeloid malignancies with inv(3)/t(3;3) reveals a predominant involvement of RAS/RTK signaling pathways, Blood*,* 125 (2015), 133-9.

3 Taskesen E, Havermans M, van Lom K, **Sanders MA**, van Norden Y, Bindels E, Hoogenboezem R, Reinders MJ, Figueroa ME, Valk PJM, Lowenberg B, Melnick A, and Delwel R, Two splice-factor mutant leukemia subgroups uncovered at the boundaries of MDS and AML using combined gene expression and DNA-methylation profiling, Blood*,* 123 (2014), 3327-35.

4 Abbas S, **Sanders MA**, Zeilemaker A, Geertsma-Kleinekoort WM, Koenders JE, Kavelaars FG, Abbas ZG, Mahamoud S, Chu IW, Hoogenboezem R, Peeters JK, van Drunen E, van Galen J, Beverloo HB, Lowenberg B, and Valk PJM, Integrated genome-wide genotyping and gene expression profiling reveals BCL11B as a putative oncogene in acute myeloid leukemia with 14q32 aberrations, Haematologica*,* 99 (2014), 848-57.

5 van der Velden VH, Hoogeveen PG, de Ridder D, Schindler-van der Struijk M, van Zelm MC, **Sanders MA**, Karsch D, Beverloo HB, Lam K, Orfao A, Lugtenburg PJ, Bottcher S, van Dongen JJ, Langerak AW, Kappers-Klunne M, and van Lom K, B-cell prolymphocytic leukemia: a specific subgroup of mantle cell lymphoma, Blood*,* 124 (2014), 412-9.

6 Gröschel S*, **Sanders MA***, Hoogenboezem R, de Wit E, Bouwman BA, Erpelinck C, van der Velden VH, Havermans M, Avellino R, van Lom K, Rombouts EJ, van Duin M, Dohner K, Beverloo HB, Bradner JE, Dohner H, Lowenberg B, Valk PJM, Bindels EM, de Laat W, and Delwel R, A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia, Cell*,* 157 (2014), 369-81.

7 **Sanders MA**, and Valk PJM, The evolving molecular genetic landscape in acute myeloid leukaemia, Curr Opin Hematol*,* 20 (2013), 79-85.

8 **Sanders MA**, and Valk PJM, Genome-wide gene expression profiling, genotyping, and copy number analyses of acute myeloid leukemia using Affymetrix GeneChips, Methods Mol Biol*,* 1015 (2013), 155-77.

9 Beekman R, Valkhof MG, **Sanders MA**, van Strien PM, Haanstra JR, Broeders L, Geertsma-Kleinekoort WM, Veerman AJ, Valk PJM, Verhaak RG, Lowenberg B, and Touw IP, Sequential gain of mutations in severe congenital neutropenia progressing to acute myeloid leukemia, Blood*,* 119 (2012), 5071-7.

10 Alemdehy MF, van Boxtel NG, de Looper HW, van den Berge IJ, **Sanders MA**, Cupedo T, Touw IP, and Erkeland SJ, Dicer1 deletion in myeloid-committed progenitors causes neutrophil dysplasia and blocks macrophage/dendritic cell development in mice, Blood, 119 (2012), 4723-30.

11 Noordermeer SM, Monteferrario D, **Sanders MA**, Bullinger L, Jansen JH, and van der Reijden BA, Improved classification of MLL-AF9-positive acute myeloid leukemia patients based on BRE and EVI1 expression, Blood, 119 (2012), 4335-7.

12 Pratcorona M, Abbas S, **Sanders MA**, Koenders JE, Kavelaars FG, Erpelinck-Verschueren CA, Zeilemakers A, Lowenberg B, and Valk PJM, Acquired mutations in ASXL1 in acute myeloid leukemia: prevalence and prognostic value, Haematologica, 97 (2012), 388-92.

13 Ribeiro AF, Pratcorona M, Erpelinck-Verschueren C, Rockova V, **Sanders MA**, Abbas S, Figueroa ME, Zeilemaker A, Melnick A, Lowenberg B, Valk PJM, and Delwel R, Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia, Blood, 119 (2012), 5824-31.

14 Noordermeer SM, **Sanders MA**, Gilissen C, Tonnissen E, van der Heijden A, Dohner K, Bullinger L, Jansen JH, Valk PJM, and van der Reijden BA, High BRE expression predicts favorable outcome in adult acute myeloid leukemia, in particular among MLL-AF9-positive patients, Blood, 118 (2011), 5613-21.

15 Taskesen E, Bullinger L, Corbacioglu A, **Sanders MA**, Erpelinck CA, Wouters BJ, van der Poel-van de Luytgaarde SC, Damm F, Krauter J, Ganser A, Schlenk RF, Lowenberg B, Delwel R, Dohner H, Valk PJM, and Dohner K, 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 (2011), 2469-75.

16 **Sanders MA**, Verhaak RG, Geertsma-Kleinekoort WM, Abbas S, Horsman S, van der Spek PJ, Lowenberg B, and Valk PJM, SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels, BMC Genomics, 9 (2008), 41.

17 Wouters BJ, **Sanders MA**, Lugthart S, Geertsma-Kleinekoort WM, van Drunen E, Beverloo HB, Lowenberg B, Valk PJM, and Delwel R, Segmental uniparental disomy as a recurrent mechanism for homozygous CEBPA mutations in acute myeloid leukemia, Leukemia, 21 (2007), 2382-4.

18 Verhaak RG, **Sanders MA**, Bijl MA, Delwel R, Horsman S, Moorhouse MJ, van der Spek PJ, Lowenberg B, and Valk PJM, HeatMapper: powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics, BMC Bioinformatics, 7 (2006), 337.

19 Verhaak RG, Goudswaard CS, van Putten W, Bijl MA, **Sanders MA**, Hugens W, Uitterlinden AG, Erpelinck CA, Delwel R, Lowenberg B, and Valk PJM, Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance, Blood, 106 (2005), 3747-54.

* These authors contributed equally to this work

APPENDIX

Abbreviations

4C-Seq	Circularized chromosome conformation capture deep-sequencing
ABL1	c-Abl oncogene 1
ALL	Acute lymphoblastic leukemia
Allo-HSCT	Allogeneic hematopoietic stem cell transplantation
AML	Acute myeloid leukemia
AML1	Acute myeloid leukemia 1
ASXL1	Additional sex combs like 1 (Drosophila)
Auto-HSCT	Autologous hematopoietic stem cell transplantation
BAC	Bacterial artificial chromosome
BCL11B	B-cell CLL/lymphoma 11B
BCR	Breakpoint cluster region
BET	Bromodomain and extraterminal domain family
BM	Bone marrow
BRD4	Bromodomain-containing protein 4
BTLA	B and T lymphocyte associated
Cas9	CRISPR associated protein 9
CBF	Core binding factor
CBFB	Core binding factor beta subunit
CD	Cluster of differentiation
CDKN2A/B	Cyclin-dependent kinase inhibitor 2A/2B
CEBPA	CCAAT/enhancer binding protein alpha
CGH	Comparative genomic hybdridization
Chip-Seq	Chromatin immunoprecipitation followed by deep-sequencing
CI	Confidence interval
CLP	Common lymphoid progenitor
CML	Chronic myeloid leukemia
CML-BC	Chronic myeloid leukemia in blast crisis
CMP	Common myeloid progenitor
CN	Cytogenetical normal
CNV	Copy number variation
CNVsvd	Copy number variation by singular value decomposition package
CR	Complete remission
CRISPR	Clustered regularly interspaced short palindromic repeat
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference

cRSS	Cryptic recombination signal sequence
CSF3	Colony-stimulating factor 3
CSF3R	Colony-stimulating factor 3 receptor
CSF3R-d715 to d730	Mutated CSF3R, truncated receptors at amino acid position 715 to 730
CSF3R-T595I	Mutated CSF3R, substitution of threonine to isoleucine at amino acid position 595
CSF3R-T595V	Mutated CSF3R, substitution of threonine to valine at amino acid position 595
CTS	Common translocated segment
dHPLC	Denaturing high performance liquid chromatography
DNA	Deoxyribonucleic acid
DNA-pkcs	DNA-dependent protein kinase catalytic subunit
DNMT	DNA methyltransferase
DSB	Double strand break
EFS	Event-free survival
ELANE	Neutrophil elastase
emPCR	emulsion PCR
ErP	Erythrocyte precursor
ETO	Eight twenty one
ETP	Early thymocyte progenitor
EVI1	Ecotropic virus integration site 1
EZH2	Enhancer of zeste homolog 2
FAB	French-American-British
FISH	Fluorescence in situ hybridization
FLT3	fms-related tyrosine kinase 3
GATA2	GATA binding protein 2
G-CSF	Granulocyte colony stimulating factor
GEO	Gene expression omnibus
GEP	Gene expression profiling
GFP	Green fluorescent protein
GM-CSF	Granulocyte-macrophage colony stimulating factor
GMP	Granulocyte monocyte progenitor
GMP	Granulocyte myeloid precursor
gRNA	Guide RNA
HDR	Homology directed repair
HMGB1	High-mobility group protein B1

HMGB2	High-mobility group protein B2
HMM	Hidden Markov model
HOVON	Dutch-Belgian Hemato-Oncology Cooperative Group
HR	Hazard ratio
HSC	Hematopoietic stem cell
HSCP	Hematopoietic stem cell and progenitor
IDH1	Isocitrate dehydrogenase 1
IDH2	Isocitrate dehydrogenase 2
IKZF1	IKAROS family zinc finger 1
IL-2/-3/-6	Interleukin 2/3/6
Indel	Insertion or deletion
IP	Immunoprecipitation
ITD	Internal tandem duplication
JAK2	Janus kinase 2
KMT2A	lysine (K)-specific methyltransferase 2A
KRAS	V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
L-BFGS-B	Bounded low-memory BFGS
LINE	Long interspersed nuclear element
LLGL2	Lethal giant larvae homolog 2 (Drosophila)
lncRNA	Long non-coding RNA
LOH	Loss-of-heterozygosity
LSC	Leukemic stem cells
LTR	Long terminal repeat
MAS	MicroArray Suite
MDS	Myelodysplastic syndrome
MECOM	MDS1 and EVI1 complex locus
MEP	Megakaryocyte erythrocyte precursor
mESC	Mouse embryonic stem cells
MkP	Megakaryocyte precursor
MLL	Mixed lineage leukemia
MPN	Myeloproliferative neoplasm
MPP	Multipotent progenitor
mRNA	Messenger RNA
MYH11	Myosine, heavy chain 11
NF1	Neurofibromin 1

NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NOTCH1	Notch homolog 1
NPM1	Nucleophosmin
NRAS	Neuroblastoma RAS viral oncogene homolog
OS	Overall survival
PAX5	Paired box 5
PB	Peripheral blood
PCR	Polymerase chain reaction
PML	Promyelocytic leukemia
PRC2	Polycomb repressive complex 2
PTD	Partial tandem duplication
RAEB(t)	Refractory anemia with excess of blasts(in transformation)
RAG1	Recombination activating gene 1
RAG2	Recombination activating gene 2
RARA	Retinoic acid receptor, alpha
RAS	Rat sarcoma
RFS	Relapse free survival
RNA	Ribonucleic acid
RNA-Seq	RNA profiling by deep-sequencing
RQ-PCR	Quantitative real-time reverse transcription PCR
RSS	Recombination signal sequence
RTK	Receptor tyrosine kinase
RT-PCR	Reverse transcription PCR
RUNX1	Runt-related transcription factor 1
SCN	Severe congenital neutropenia
SF3B1	Splicing factor 3B subunit 1
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
STAT3/5	Signal transducer and activator of transcription
SUZ12	Supressor of zeste polycomb repressive complex 2 subunit
SV	Structural variant
TAD	Transactivation domain
TALEN	Transcription activator-like effector nuclease

TCGA	The cancer genome atlas
TET2	Tet methylcytosine dioxygenase 2
TKD	Tyrosine kinase domain
UPD	Uniparental disomy
VAF	Variant allele frequency
WBC	White blood cell count
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World health organization
WT1	Wilms tumor 1
ZC3H18	Zinc finger CCCH-type containing 18

APPENDIX

PhD portfolio

Name PhD student:	Mathijs A. Sanders	**PhD period:**	December 2009 - December 2014
Erasmus MC Department:	Hematology	**Promotors:**	Prof. dr. B. Löwenberg, Prof. dr. R. Delwel
Research school:	Molecular Medicine	**Supervisor:**	Dr. P.J.M. Valk

1. PhD training	Year	ECTS
General courses		
Basic course R	2009	1.4
In-depth courses		
Course on Molecular Diagnostics	2010	1
Next Generation Sequencing Training: CLC Bio	2011	0.5
Complete genomics	2011	0.3
Course on Molecular Aspects of Hematological Disorders	2014	0.5
Seminars and workshops		
Scientific workshop Acute Myeloid Leukemia "Molecular"	2011	1
Workshop on competing risk models	2011	0.5
Statistical Methods for Post Genomic Data Workshop (Paris)	2011	1
Erasmus Hematology Lectures	2009-2014	4
Oral presentations		
Workdiscussion (department of Hematology, Erasmus MC, 15x)	2009-2014	7.5
Workdiscussion (department of Medical Statistics and Bioinformatics, LUMC, 4x)	2009-2013	2
AIO/post-doc meeting (department of Hematology, Erasmus MC, 4x)	2009-2014	2
Journal club (department of Hematology, Erasmus MC, 4x)	2009-2014	2
3rd Channel Network Conference (Bordeaux Segalen university)	2011	1
Dutch hematology congress (3x)	2011-2013	3
High-dimensional data modeling (department of Medical Statistics and Bioinformatics, LUMC, 1x)	2012	0.5
Annual MODHEM/SKLM Spring Meeting	2012	1
Daniel den Hoed day (2x)	2012-2013	2
Course on Molecular Aspects of Hematological Disorders	2014	1
Annual conference American Society of Hematology (ASH)	2014	1
National and international conferences		
Benelux Bioinformatics Conference (BBC)	2009	1
Molecular Medicine day (2x)	2010-2011	0.6

3rd Channel Network Conference (Bordeaux Segalen university)	2011	1
Dutch hematology congress (4x)	2011-2014	1.8
Annual conference Center for Translational Molecular Medicine (CTMM) (3x)	2010-2012	1
Annual conference European Hematology Association (EHA) (2x)	2012, 2014	2
Annual conference American Society of Hematology (ASH) (2x)	2012, 2014	2
Scientific meetings		
Workdiscussion (department of Hematology, Erasmus MC)	2009-2014	5
AIO/post-doc meeting (department of Hematology, Erasmus MC)	2009-2014	2
Journal club (department of Hematology, Erasmus MC)	2009-2014	3.75
Workdiscussion (LUMC)	2009-2012	3
Other		
Writing grant Complete Genomics (granted)	2012	0.5
2. Teaching activities	**Year**	**ECTS**
Supervising students		
Bachelor student ("Bioinformatics" and "Molecular biology") (2x)	2011-2012	3
Two bachelor exchange students Moscow State University (LUMC)	2011	1
Supervising practical training and excursions		
Organization and supervision invited speaker lunch (2x)	2013-2014	0.2
Introduction SPSS and mixed models, Medical and Biomedical Science students (LUMC)	2010-2012	1.5
Biomedical Research Techniques (Molmed, oral presentations)	2013-2014	2
Total		**64.55**

Stellingen behorende bij het proefschrift

Computational Biology-Driven Genomic and Epigenomic Delineation of Acute Myeloid Leukemia

1. De aanduiding van het RPN1-EVI1 leukemietype moet binnen het World Health Organisation classificatiesysteem worden gewijzigd naar GATA2-EVI1 (dit proefschrift).
2. Naast het Burkitt's lymfoom en het multipel myeloom, moet ook acute myeloïde leukemie worden beschouwd als een hematologische maligniteit die gedreven kan worden door de aberrante activatie van proto-oncogenen door het herpositioneren van super-enhancers (dit proefschrift).
3. Genetische afwijkingen bij acute lymfatische leukemie met kinase-activerende afwijkingen zijn het gevolg van onrechtmatige genetische herschikkingen door het RAG eiwitcomplex (dit proefschrift).
4. Verworven mutaties in *RAS*/RTK genen dragen essentieel bij aan de leukemische transformatie van AML met inv(3)/t(3;3) afwijkingen (dit proefschrift).
5. Het combineren van verschillende genoom-brede technieken kan nieuwe mechanistische inzichten verschaffen in de ontwikkeling van leukemie (dit proefschrift).
6. Mutaties in DNA methylatie-gerelateerde genen in pre-leukemische stamcellen spelen een belangrijke rol in het vroege ontstaan van acute myeloïde leukemie. (Shlush et al, Nature 2014;506:328-33; Jaiswal et al, NEJM 2014;371:2488-98; Genovese et al, NEJM 2014;371:2477-87)
7. De functionele synergie tussen verschillende co-existente genmutaties is als een complex raderwerk; pas wanneer de functie van elk onderdeel bekend is, kan door kleine aanpassingen in de functie van elk van de onderdelen het mechaniek van het geheel begrepen worden.
8. Communicatieve vaardigheden zijn essentieel bij de uitvoering van multidisciplinair onderzoek.
9. Falsificationisme is de grootste vijand van menig wetenschapper.
10. Geen één leukemie is hetzelfde.
11. "When you are studying any matter, or considering any philosophy, ask yourself only what are the facts and what is the truth that the facts bear out. Never let yourself be diverted either by what you wish to believe, or by what you think would have beneficent social effects if it were believed, but look only, and solely, at what are the facts." (*Bertrand Russell*)

Mathijs Sanders
Rotterdam, 17 februari 2015