Computational Biology in Acute Myeloid Leukemia with *CEBPA* Abnormalities

Erdoğan Taşkesen

PROMOTIECOMMISSIE

Promotor:
Prof.dr. H.R. Delwel

Overige leden:
Prof.dr. B. Löwenberg
Prof.dr.ir M.J.T. Reinders
Prof.dr. J.N.J. Philipsen

Table of Contents

CHAPTER 1

General Introduction

General Introduction

SUMMARY

In the last decade, tiling-array and next-generation sequencing technologies allowed quantitative measurements of different cellular processes, such as mRNA expression, genomic changes including deletions or amplifications, DNA-methylation, chromatin modifications or Protein-DNA-binding interactions. Using these technologies, thousands of features can now be measured simultaneously in a patient cell sample. The use of for instance mRNA expression profiles or DNA-methylation profiles have already provided new insight into the molecular biology of patients with Acute Myeloid Leukemia (AML). AML is a blood cell malignancy, in which primitive myeloid cells have been transformed and accumulate in the bone marrow and blood. Different forms of AML exist with different molecular abnormalities that associate with distinct responses to therapy. Many subgroups with comparable mRNA expression or DNA-methylation patterns were identified. These studies also revealed the existence of novel previously undefined AML subtypes. Among those was a group of patients with a mutation in a gene called *CEBPA*. *CEBPA* is a gene that encodes the transcription factor CCAAT Enhancer Binding Protein Alpha (C/EBPα), which controls the expression of genes in myeloid progenitor cells. Mutated *CEBPA* encodes a dysfunctional C/EBPα-protein, which consequently results in aberrant control of "target genes". In this thesis we focus particularly on the role of *CEBPA*. We studied the predictive and prognostic relevance of mutated *CEBPA*, and analyzed in a genome wide fashion the mRNA expression, DNA-methylation and the protein-DNA-binding levels corresponding to (mutated) *CEBPA* in AML. For the analysis of protein-DNA-binding, we developed a novel statistical methodology. With this statistical methodology we studied the fundamental role of (mutant) C/EBPα binding and the effect on gene expression levels. We also integrated gene expression with DNA-methylation profiles of hundreds of AML patients and revealed the existence of two previously unidentified AML subtypes.

FROM HEALTHY TO CANCEROUS CELLS

Cells in a living organism are designated with a functional role that can be classified by their tissue of origin. In human, there are many distinct cell types[1,2] for which their function varies widely. In general, a cell (Figure 1A) contains a nucleus (Figure 1B) that contains the genetic instructions in the chromosomes which are essential for the development and functioning of a particular cell. In humans we count 46 individual chromosomes or 23 chromosome pairs (Figure 1C), which consist of DeoxyriboNucleic Acid (DNA). Stretches of the DNA with specific functions are denoted as "genes"[3,4] (Figure 1D). In order to get a healthy functioning cell, genes need to be transcribed to messenger RiboNucleic Acid (mRNA, Figure 1E) and then translated into proteins[5] (Figure 1F). The proteins are the functional units of the cell. In the human genome, approximately 23000 protein-coding genes are identified[6]. The function of a particular cell depends on the combination of genes that are in "onset" or "offset". The communication between

genes and proteins is part of a complex system and errors in these crucial processes are responsible for diseases such as cancer. A majority of genes are responsible for the maintenance of basic cellular functions[7,8]. Other genes encode for proteins that are needed for specific functions of a particular cell[1]. When DNA is damaged at the gene-locations, genes may become malfunctioning or non-functioning. When this happens, cancer may arise. Cancer is a generic term for a large group of diseases that have different incidence and mortality rates[9].

The one defining feature of cancer is that cells in tumorigenic state grow too fast and beyond their usual boundaries that may consequently result in an invasion to other body parts and organs, which is called metastasis[10]. Such a tumorigenic state can only be entered when existing fail-safe mechanisms are bypassed or are shutdown[11], which is often due to genetic changes in genes that are controlling these mechanisms[12]. Two classes of genes can be discriminated that cause far reaching effects when damaged: oncogenes[13] and tumor suppressor genes[14]. These genes do often have a role in the fail-safe mechanisms of a cell. Oncogenes will drive a cell towards tumorigenic activity once hyper-activated[13]. Conversely, shutting down a tumor suppressor gene may lead to increased tumorigenic activity as well[14].

Leukemia is such a cancerous disease. As in other forms of cancer, mutations are observed in regulatory genes (in leukemia cells). Most mutations that cause cancer are of somatic origin[15], meaning that the DNA damage is acquired, most likely as the result of environmental or endogenous carcinogenic agents. Mutations may sometimes be inherited or can be gained in a situation where individuals have increased susceptibility to develop cancer. Both somatic and germline mutations can result into abnormal cell growth and development.

BLOOD CELL FORMATION AND ACUTE MYELOID LEUKEMIA

Blood cells arise from hematopoietic stem cells (HSCs), which reside in the bone marrow[16]. In a normal (healthy) situation, HSCs develops into primitive progenitor cells which may then mature into red blood cells (erythrocytes), platelets (thrombocytes) or different types of white blood cells (leukocytes) (Figure 2)[16,17]**.** However, mutations in genes can cause that HSCs or primitive lineage specific progenitor cells are unable to develop and remain immature. These immature cells accumulate in the blood and bone marrow causing less room for healthy blood cells. Such a cancerous process is called leukemia.

Acute Myeloid Leukemia (AML) is the most common myeloid disorder in adults[9], and is the disease that is central in this thesis. The prevalence is on average 3.5 cases per 100.000 individuals world-wide but increases with age[9]. The median age of presentation is approximately 67 years and the disease is known to be heterogeneous[18], meaning that different forms of AML exist. As an example, there are groups of patients with chromosomal abnormalities, such as inversions of DNA-fragments (inv(16)) or the rearrangements of chromosomal parts that join two other separated genes, denoted as translocations (t(15;17), t(8;21)). Frequently such inversions and translocations cause gene fusions to occur, creating hybrid proteins. In other translocations, it has been suggested that a strong promoter of one gene

is repositioned to the coding sequence of another, resulting in overexpression of that particular gene. Other known abnormalities in AML are subtle mutations in cancer-critical genes, such as for Internal Tandem Duplications of FLT3 (*FLT3ITD*) [19-23], nucleophosmin (*NPM1*)[24,25] or "*CCAAT enhancer binding protein alpha*" (*CEBPA)* [26-28] (Figure 3A). These and many other abnormalities are nowadays used in the risk-classification of patients, i.e. to predict favorable, intermediate or poor treatment outcome[29-32].

The risk-classification of AML was until the late seventy's based on the pathology and cytological examination of bone marrow and blood cells[33]. Nowadays, various subtypes of AML are known, each with different survival rates[33]. Classification of the disease is important to refine more "personalized" therapy, i.e. predicting which therapy may work the best for a (group of) patient(s), and thereby increase the chance to survive. However, not all AMLs can be classified into the known AML subtypes. To characterize the underlying abnormalities of these unclassifiable AMLs we first need to detect which DNA changes (i.e. mutations) and which mechanisms or pathways were involved.

AML PATIENTS WITH ABNORMAL *CEBPA* EXPRESSION

CEBPA abnormalities are investigated for more than a decade[27] but only since the emerging microarray technology[18,34], the possibilities for fast mutational screening[35], and the use of large databanks of patient's cohorts, fast refinement and many new insights are provided. This section introduces 1. The discovery of *CEBPA* mutations, 2. The survival and treatment response of AML patients with *CEBPA* mutations, 3. The use of gene expression profiles to predict AML with *CEBPA* mutations, and 4. The interactions of C/EBPα to the DNA.

It has been shown that *CEBPA* mutations occur mainly in cytogenetically normal AML (CN-AML) with an incidence of 5-14%[27,28,31,36-42]. Two main types of *CEBPA* mutations can be distinguished: N-terminal frame-shift mutations resulting in the translation of a 30-kDa protein only, and the C-terminal in-frame mutations in the basic zipper region affecting DNA-binding and homodimerization and heterodimerization[28,43,44] (Figure 3A). It is known that *CEBPA* mutations can roughly be separated into two subgroups, i.e., AML patients with a single mutation (*CEBPA*^{sm}) and those with double mutations (*CEBPA*^{dm})[35,45-48]. Favourable outcome is observed in AML with *CEBPA*^{dm} but not for *CEBPA*^{sm} [35,44], however it is unclear why the clinical outcome between *CEBPA*^{dm} and *CEBPA*^{sm} is different. In addition, the distribution of distinct *CEBPA* mutations, and presence or absence of the concurrent mutations in other genes for the same AMLs had not been established yet. We therefore investigating the distribution of *CEBPA* mutations in a large cohort of patients (Chapter 7). We show that in the majority of *CEBPA*^{dm} AML, both alleles are mutated (Figure 3B). These biallelic mutations frequently harbour an N-terminal mutation on one allele and a C-terminal bZIP mutation on the other. In *CEBPA*^{sm} AML, mutations occur mostly in the N-terminus, although single C-terminal mutants have been found as well (Figure 3B). We furthermore analysed in Chapter 7 the presence of concurrent mutations in *CEBPA*^{dm} or *CEBPA*^{sm}. We detected that these were significantly enriched for *CEBPA*^{sm} AMLs. With this knowledge we could address why *CEBPA*^{dm} differed in clinical outcome compared to *CEBPA*^{sm}, and whether the outcome is affected by concurrent mutations.

The prognosis for AMLs with a *CEBPA*^{dm} is defined as favourable; 5-years overall survival is ranging between 50% and 70%[35,44-48]. These patients were consolidated with different types of therapy, such as intensive chemotherapy, autologous or allogeneic hematopoietic stem cell transplantation (autoHSCT or alloHSCT). After such treatment protocols, patients may enter a so called complete remission (CR), meaning that less than 5 percent blast cells are present in the bone marrow, and none can have a leukemic phenotype[49,50]. Relapse remains a major cause of treatment failure and occurs frequently within the first 2 years after entering a complete remission (CR). This has for instance raised the question whether a HSCT in first CR should be recommended in patients with this form of AML. Analyses according to type of post remission treatment in the subset of AML with mutant *CEBPA*, for clinical reasons useful, have so far not become available mainly due to limited patient numbers precluding meaningful statistical analyses. In Chapter 8 we studied AMLs *CEBPA* mutational status (age 18-60 years) in AML. The benefit of post remission stem cell transplantation in AML patients that harbour *CEBPA*^{dm} compared to *CEBPA*^{dm} AMLs which were not treated with stem cell transplantation was investigated.

It has previously been discovered that *CEBPA*^{dm} has a uniquely associated gene expression profile[35] which stresses the notion to mark *CEBPA*^{dm} AMLs as a distinctive disease entity. This means that these AMLs have very similar gene expression profiles which can subsequently be used for diagnostic purposes[51], such as prediction of *CEBPA*^{dm} AMLs given the gene expression profiles. A predictive gene signature[35] has therefore been created but was hampered by AMLs with hypermethylation of the proximal promoter region of *CEBPA*[34]. In addition, it is unknown whether classification results are affected by homozygous N-terminal or C-terminal *CEBPA*^{dm} mutations or because of germline mutations. In Chapter 7 we created a gene signature that has increased the power for *CEBPA*^{dm} prediction, and we addressed the question whether homozygous N-terminal or C-terminal and germline *CEBPA*^{dm} mutations showed differences in the classification.

Although a very specific gene signature is created that describes *CEBPA*^{dm} AMLs, it does not describe what the functional effect is of mutated *CEBPA* in primary AMLs. It is known that the N-terminal domain contributes to cell growth inhibition, whereas the C-Terminal domain that contains a basic region required for DNA-binding and a leucine zipper (bZIP) essential for homo and heterodimerization[44,52-54]. However, the functional effect of mutated *CEBPA* in primary AML cells is unknown. We studied the binding capacity of a mutated C-terminal *CEBPA* in Chapter 5 and provide data that suggest that C-terminal mutant *CEBPA* is capable of (in) direct binding to the DNA.

DETECTION OF FUNCTIONAL REGIONS IN THE GENOME

In the last decade, technology such as tiling-array hybridizations provided whole genome coverage by using probes and therefore useful for exploring the genome in an unbiased fashion. Tiling-array technology is valuable for different applications, such as 1. Protein-DNA-interaction by conducting chromatin immunoprecipitation followed by array (Chip) hybridization (ChIP-on-chip, Figure 1H) experiments, 2. Epigenetic modifications by Methyl-DNA immunoprecipitation (MeDIP-on-chip, Figure 1G) or 3. Identification of DNAse hypersensitive sites, which can be used

to predict regulatory elements such as promoter regions, enhancers and silencers. Each tiling-array produces quantitative signal intensity for each probe by the hybridization of labeled DNA. Probe intensities are illustrated by the different peaks in Figure 1H. Single probe-hybridization with high signal intensity suggests strong hybridization but it is not necessarily the result of specific hybridization of labeled DNA. Multiple contiguous probes that show increased signal intensity upon hybridization across a particular genomic region are more likely to be the result of true hybridization in a biological experiment. To detect biologically relevant genomic regions, probe intensity signals should be discriminated from non-specific signals. A challenge in the analysis of tiling-array data is finding those genomic regions where the signal significantly deviates from the general genome wide behavior. Determining these candidate regions is difficult as there is no "*golden standard*" that defines properties (e.g. signal intensity or size) of such region. We therefore developed a novel statistical methodology, Hypergeometric Analysis of Tiling-arrays (HAT), (Chapter 2) to identify candidate genomic regions in tiling-array data. The use of a dynamic window makes our model independent of size and therefore applicable for different biological tiling-array experiments (protein-DNA-binding, DNA-methylation, histone modifications). We used HAT to study the binding of mutated C/EBPα (Chapter 5), wild-type C/EBPα (Chapter 6), and to detect viral integration sites that potentially harbour new tumour suppressor genes (Chapter 4).

HAT showed to be successful[55-57] for the analysis of tiling-array data. However, in the past few years Next-Generation Sequencing technology (NGS) has rapidly replaced tiling-array hybridization because of the increased resolution with which the interactions can be measured. Instead of using probes, the immunoprecipitated DNA-fragments in a ChIP experiment are sequenced and in turn aligned to the reference genome. We developed (Chapter 3) HATSEQ (Hypergeometric Analysis of Tiling-arrays and Sequence data), which is an extremely scaled version of HAT that can work on a base resolution, and has proven to be accurate in the detection of potential candidate regions.

DETECTION OF AML SUBTYPES BY USING GENE EXPRESSION AND DNA-METHYLATION PROFILES

Ideally, when it comes to cancer, we want to reconstruct all the changes that occurred for a single cancer to determine exactly where, what and how it went wrong in the DNA. This requires measuring various different processes, such as DNA-methylation[58] (Figure 1G) or mRNA expression levels[18] (Figure 1E). Although this is possible, the analysis of large data sets is not a routine process. In our simplified scheme (Figure 1) we show with traffic-lights that DNA-methylation can be seen as a logical "*AND*" operator for the onset or offset of genes (Figure 1D). However, this process is not deterministic but rather a stochastic process, and may depend on other factors (such protein-DNA-binding, Figure 1H), and therefore difficult for routine analysis. These processes (Figure 1G, H, E and F) are so far investigated thoroughly but mostly independently from each other[18,58-60]. Gene expression profiling (GEP) has been used to analyse the mRNA gene expression profiles of hundreds of AML patients[18]. For the same patients, DNA-methylation profiles (DMP) are also measured[58]. Both technologies provided multiple data sets which could lead to novel insights into

AML[34,61,62], such as the discovery of novel AML subgroups. We hypothesized that the combined data results in specific patterns in cancer cells that may uncover novel AML subgroups. In Chapter 9 we developed an approach to combine the two data sets and we identified two novel AML subgroups that could not be identified using GEP[18] or DMP[58] alone.

SCOPE AND AIM OF THE THESIS

The thesis presents work divided into four sections. The first section (Chapter 2, 3 and 4) deals with the detection of candidate regions in the genome using tiling-array and next-generation sequencing technology. These technologies have been developed to accurately determine potential functional regions in the genome, such as for the binding of proteins that regulate transcription. Chapter 2 describes the statistical method that we developed to detect candidate regions for tiling-array technology. Although the many successes of tiling-array technology, next-generation sequencing technology rapidly replaces tiling-arrays because of the increased resolution with which the interactions can be measured. In Chapter 3 we demonstrate how we optimized our methodology to work on a base resolution, and to detect potential candidate regions in the genome. This Chapter is followed by a case study where viral integration sites are identified that potentially harbour new tumour suppressor genes in a so called MeDIP-on-chip dataset (Chapter 4).

The second section (Chapter 5 and 6) of the thesis is concerned with the experimental interactions of C/EBPα in an inducible myeloid cell line model. In Chapter 5 we used our novel methodology, and asked the question whether we could identify potential targets of mutated C/EBPα. In AML patients it has previously been shown that silencing of *CEBPA* leads to cells with myeloid/T-lymphoid features. However, the exact role of C/EBPα in this process is unknown. In Chapter 6 we asked the question whether the newly developed computational technologies can be of use to study the mechanism by which silencing of *CEBPA* may play a role in transformation of cells with myeloid/T-lymphoid characteristics.

The third section of the thesis (Chapter 7 and 8) focuses on AML patients with mutations in *CEBPA*. It is known that mutations in *CEBPA* occur in a biallelic (double) or mono allelic (single) fashion. AML patients with *CEBPA* double mutations are associated with favourable outcome whereas patients with single mutations in *CEBPA* showed unfavourable outcome. In Chapter 7 we evaluated the outcome of *CEBPA* double and single mutations with respect to other concurrent mutations. In Chapter 8 we asked the question whether the favourable prognosis of AML with *CEBPA* double mutation is to be attributed to a distinct post remission strategy, i.e. treatment of allogeneic or autologous hemapoetic stem cell transplantation (alloHSCT, autoHSCT respectively), compared to patients treated with chemotherapy.

The fourth, and final section (Chapter 9) centres on the questions whether the combined gene expression profiles and DNA-methylation profiles can be used to identify previously unrecognized subgroups of AML.

Chapter 10 provides a general discussion of the results and future perspectives are provided.

FIGURE LEGENDS

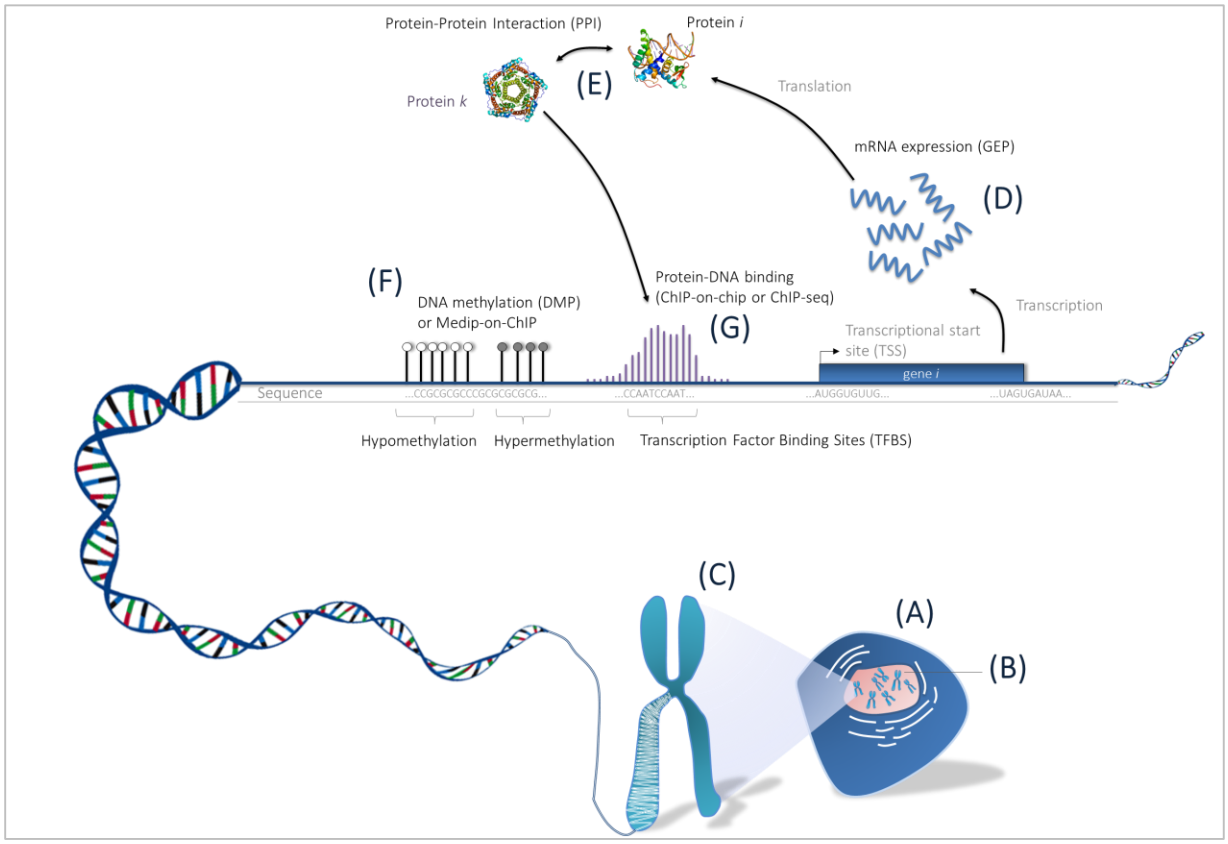


Figure 1. Overview of large scale measurements. The cell (A) which contains a nucleus (B), which in turn contains the chromosomes (C) containing all the genetic information. The following data are frequently measured large scale: mRNA expression by using Gene Expression Profiling (GEP) (D), Protein-protein interactions (PPI) (E), DNA-methylation levels (DMP or MeDIP-on-chip) (F), or protein-DNA-binding interactions (ChIP-on-chip or ChIP-Seq) (G). DNA helix is adapted from National Human Genome Research Institute.

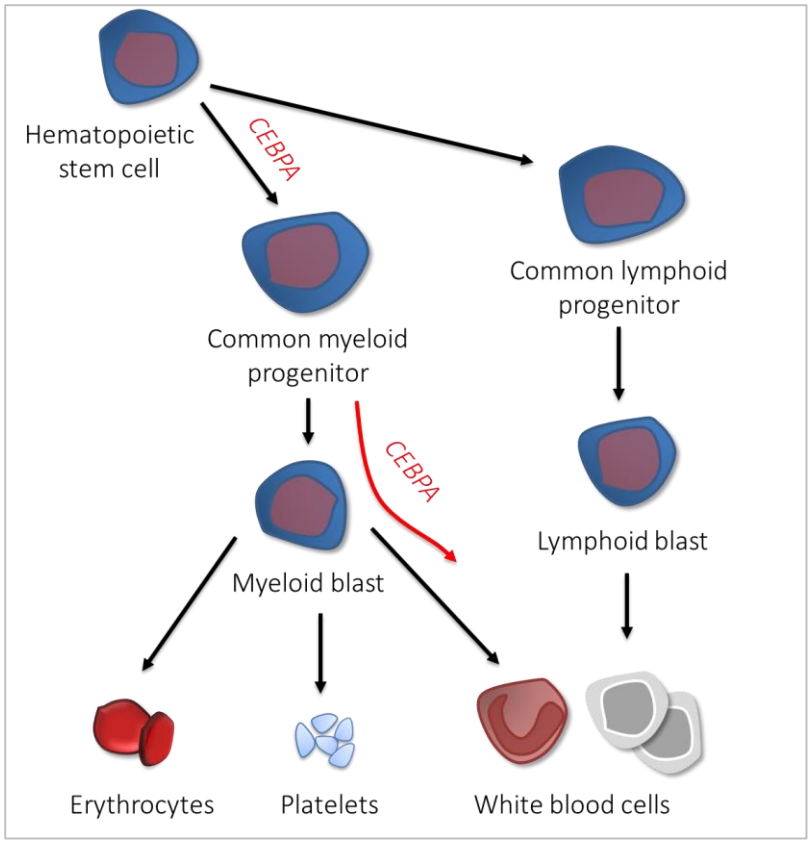


Figure 2. Overview of hematopoietic stem cell development. A hematopoietic stem cell matures into myeloid blasts or lymphoid blasts. The myeloid blast can subsequently mature into the Erythrocytes, Platelets or various types of white blood cells. A lymphoid blast matures into white blood cell such as, B-cells or T-cells[63].

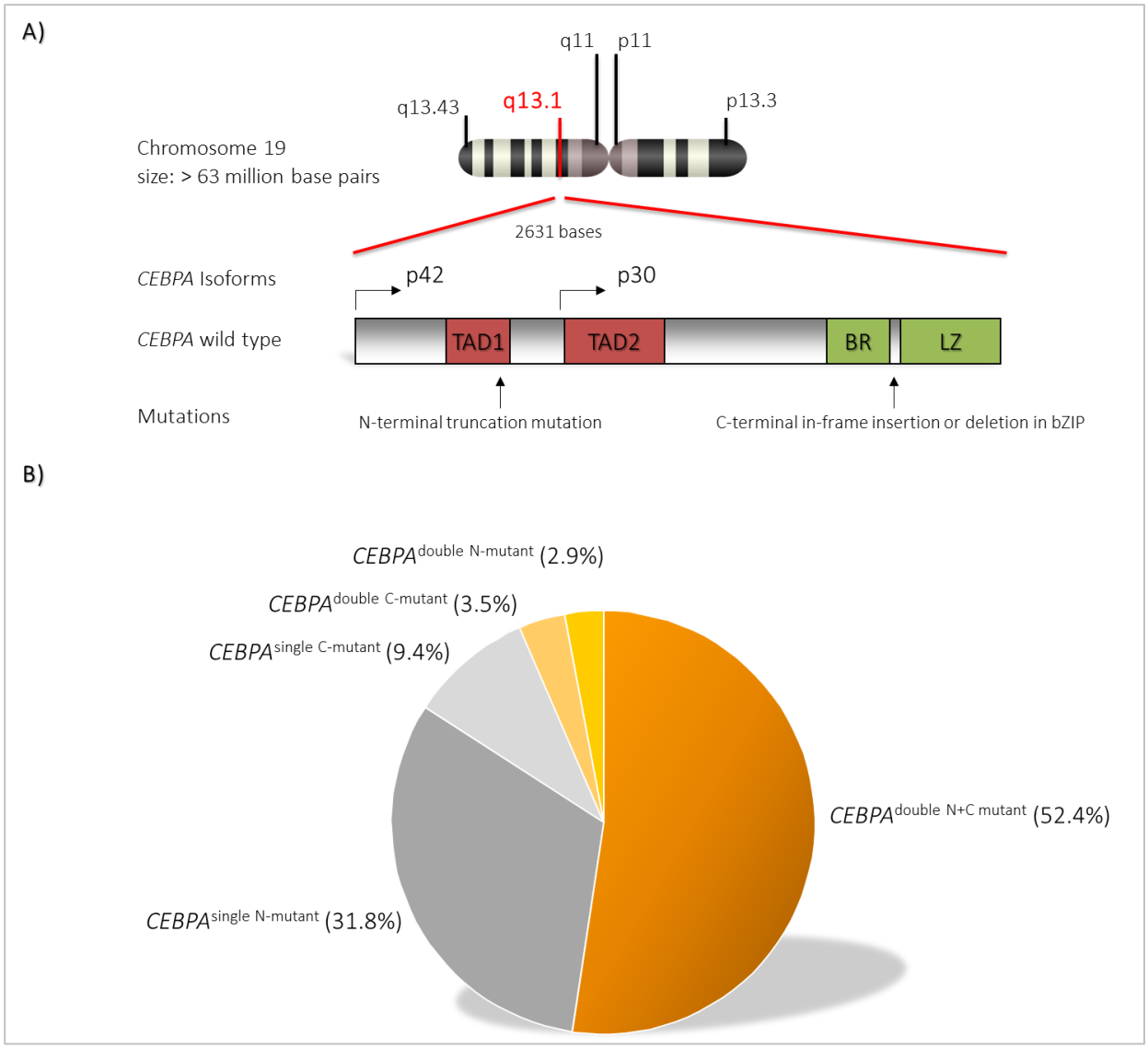


Figure 3. Overview of *CEBPA*. (A) Gene *CEBPA* lies at chromosome 19. There are two isoforms that are translated from the same mRNA: P42 and P30. Mutations in *CEBPA* are seen in the N-terminal and C-terminal regions. (B) The frequency of mutated *CEBPA* (N/C-terminal) for 170 AML patients (data is used from HOVON-SAKK and AMLSG-cohort[44]).

CHAPTER 2

HAT: Hypergeometric Analysis of Tiling-arrays with Application to Promoter-GeneChip Data

BMC Bioinformatics

May 2010 | Volume 11 | Issue 1 | doi: 10.1186/1471-2105-11-275

HAT: Hypergeometric Analysis of Tiling-arrays with Application to Promoter-GeneChip Data

Erdogan Taskesen, Reneé Beekman, Jeroen de Ridder, Bas J. Wouters, Justine K. Peeters, Ivo P. Touw, Marcel J.T. Reinders, Ruud Delwel

ABSTRACT

Background: Tiling-arrays are applicable to multiple types of biological research questions. Due to its advantages (high sensitivity, resolution, unbiased), the technology is often employed in genome wide investigations. A major challenge in the analysis of tiling-array data is to define regions-of-interest, i.e., contiguous probes with increased signal intensity (as a result of hybridization of labeled DNA) in a region. Currently, no standard criteria are available to define these regions-of-interest as there is no single probe intensity cut-off level, different regions-of-interest can contain various numbers of probes, and can vary in genomic width. Furthermore, the chromosomal distance between neighboring probes can vary across the genome among different arrays.

Results: We have developed Hypergeometric Analysis of Tiling-arrays (HAT), and first evaluated its performance for tiling-array datasets from a Chromatin Immunoprecipitation study on chip (ChIP-on-chip) for the identification of genome wide DNA-binding profiles of transcription factor C/EBPα (used for method comparison). Using this assay, we can refine the detection of regions-of-interest by illustrating that regions detected by HAT are more highly enriched for expected motifs in comparison with an alternative detection method (MAT). Subsequently, data from a retroviral insertional mutagenesis screen were used to examine the performance of HAT among different applications of tiling-array datasets. In both studies, detected regions-of-interest have been validated with (q)PCR.

Conclusions: We demonstrate that HAT has increased specificity for analysis of tiling-array data in comparison with the alternative method, and that it accurately detects regions-of-interest in two different applications of tiling-arrays. HAT has several advantages over previous methods: i) as there is no single cut-off level for probe intensity, HAT can detect regions-of-interest at various thresholds, ii) it can detect regions-of-interest of any size, iii) it is independent of probe-resolution across the genome, and across tiling-array platforms and iv) it employs a single user defined parameter: the significance level. Regions-of-interest are detected by computing the Hypergeometric probability, while controlling the Family Wise Error. Furthermore, the method does not require experimental replicates, common regions-of-interest are indicated, a sequence of interest can be examined for every detected region-of-interest, and flanking genes can be reported.

BACKGROUND

Tiling-arrays are used for the identification of specific genomic DNA regions that can be enriched using various procedures to study certain molecular biological features. For example, DNA-fragments that are bound by a protein-of-interest, e.g., a transcription factor, can be enriched by using Chromatin Immunoprecipitation (ChIP). When these enriched fragments are hybridized to an array, a genome wide protein binding profile can be obtained that is associated with this particular protein-of-interest in the cell type that was studied (ChIP-on-chip[64]). Other applications of tiling-arrays[65] are: Methylated DNA immunoprecipitation (MeDIP-on-chip[66]), transcriptome mapping[67], recognition of hypersensitive sites such as segments of open chromatin that are cleaved more readily by DNaseI (DNase-chip[68]), or identification of copy number variations or breakpoints (Array CGH[69]). The use of tiling-arrays to detect enriched DNA regions has several advantages such as *i)* high sensitivity, which allows the detection of small DNA-fragments associating with rare molecules and, *ii)* high probe-resolution, which results in accurate acquisition of unbiased data.

A tiling-array is an array of short DNA-fragments, which represent 'probes' that cover the entire genome, or contigs of the genome. The hybridization of labeled DNA to an array (for example DNA enriched using ChIP), will produce a quantitative signal intensity for each probe. Multiple contiguous probes with increased signal intensity across a particular genomic region, is a putative region-of-interest, and suggests the presence of a protein binding site.

As there are no standard criteria to accurately define a region-of-interest, a major challenge in the analysis of tiling-array data is to define such a region, and discriminate a positive signal from non-specific signals[70]. Defining regions-of-interest requires intensity thresholds on continuous probe intensity levels. Following this, the decision of the number of consecutive probes above the threshold needs to be made before a region-of-interest is called. This threshold, and the number of probes above the threshold, directly influence the size of the region-of-interest that can be detected. As biologically relevant regions may vary in intensity, employing a single threshold is insufficient. Additionally, as the probe-resolution varies across the genome, and across different tiling-array platforms, choosing a fixed number of consecutive probes as a region-of-interest is also inadequate. Various methods have been developed to detect regions-of-interest in ChIP-on-chip data such as Welch t-test, HMM, TileMap, MAT, Mixture model approach, CMARRT, Starr and Ringo[71 72-78]. MAT (Model-based analysis of tiling-arrays for ChIP-on-chip)[74] is one of the most cited methods for analyzing ChIP-on-chip data and it has been shown to outperform Welch t-test, HMM and TileMap[71-73]. MAT uses various user defined parameters to model a region-of-interest, such as maximum bandwidth, maximum gap size between probes, the minimum number of probes in a region, and the use of a fixed threshold. A major limitation of this method is that it assumes a uniform probe-resolution across the genome, and depends on many user defined parameters.

Here, we propose a statistical framework (HAT: Hypergeometric Analysis of Tiling-arrays) to identify regions of-interest in tiling-array data. HAT has several advantages over previous methods including MAT: *i)* as there is no single cut-off level for probe intensity, HAT can detect regions-of-interest for a large number of thresholds, *ii)* it can detect regions-of-interest of any size, *iii)* it is independent of probe-resolution across the genome and across tiling-array platforms

and *iv)* it employs only a single user defined parameter: the significance level. HAT can be seen as a generalization of the transcript discovery approach used in Bertone *et al* [67].

A detailed description of our framework (Figure 1) can be found in the method section. Briefly, instead of a single probe intensity cut-off level, HAT evaluates a large number of thresholds. Each threshold transforms the continuous signal intensity levels into discrete calls for each probe; referred to as positive probes where the probe intensity exceeds the threshold, and negative probes where it does not. In order to define regions-of-interest, all probes within the window of each positive probe are evaluated and the *P-value* is defined based on the ratio of both positive and negative probes using the Hypergeometric distribution. To detect regions-of-interest of any size, the width of the window is also varied across all relevant window widths, where a relevant window is defined by the expected fragment size in the experimental procedure (e.g., due to sonication). The resulting regions-of-interest for each setting of the threshold and each window width are combined by taking the union of the significant window positions. The Family Wise Error (FWE) is controlled by employing a Bonferroni correction.

We have used two datasets using promoter tiling-arrays to evaluate HAT. In the first assay, tiling-array data was employed to identify genome wide DNA-binding profiles of the transcription factor C/EBPα, in a cell line model. Using these data, we have shown that although HAT detected fewer regions-of-interest than MAT, the detected regions are more highly enriched for CEBP binding motifs, and include known C/EBPα target genes. In the second experiment, a retroviral insertional mutagenesis assay, HAT identified novel putative transforming loci that may play a role in tumor development. Two of these loci were subsequently validated using PCR.

HAT can also detect and compare regions-of-interest across multiple samples. Each sample is analyzed independently, but when multiple samples within one experiment are used, detected regions-of-interest at the same genomic location among different samples are combined into 'common regions-of-interest', thereby increasing the confidence. In addition, HAT can incorporate sequence information for the detection of pre-defined sequences (e.g., binding location within or near the region). These are highlighted in the graphical output for every detected region-of-interest and indicated in the output file.

RESULTS AND DISCUSSION

Data

Two distinct experimental datasets were used in this study: ChIP-on-chip data derived from an inducible *CEBPA* expressing myeloid cell line model and data obtained from genomic DNA from retrovirus induced murine leukemias. Data were generated using the Affymetrix GeneChip Mouse Promoter 1.0 Array. This chip generates 4.6 million perfect match probes over 28000 mouse promoter regions. Promoter regions cover 6 Kb upstream to 2.5 Kb downstream of 5' transcription start sites. Each probe has a size of 25nt.

Detection of regions-of-interest for C/EBPα chromatin immunoprecipitation by applying HAT

To compare different methods and to analyze the promoter array data, we made use of a dataset that was obtained from a ChIP of beta-estradiol induced C/EBPα in a myeloid cell line, 32D, followed by promoter array hybridizations. The data were used to examine the validity of detected regions-of-interest in two ways: *i)* at the 'CCAAT' binding level; C/EBPα interacts with the nucleotide sequence 'CCAAT' within the promoter regions represented on the chip, therefore CEBP binding motifs are expected to be enriched, and *ii)* at the gene level; examination of the presence of known C/EBPα target genes, by taking the genes flanking the detected region-of-interest into account. Furthermore, one selected region-of-interest was validated by Real Time Quantitative PCR (qPCR).

The experimental setup was as follows: clones were derived from a myeloid cell line model (32D), that expresses either beta-estradiol inducible C/EBPα-ER (3 clones) or control-ER (2 clones). Chromatin immunoprecipitations were carried out using an antibody directed against ER in the beta-estradiol treated cells and the DNA obtained from these cells, after immunoprecipitation, was hybridized to Affymetrix promoter chips.

For method comparison we used Model-based analysis of tiling-arrays for ChIP-on-chip (MAT), with the default parameters for the detection of regions-of-interest (bandwidth of 300bp; resulting in 2*bandwidth probe positions, 300bp of maximum gap size between positive probes, minimum of 8 probes for MAT-score, and enriched fragments at the 1 × 10-5 significance level). The default settings agree with the average sonicated fragment sizes, being 600bp, and the distance between two consecutive probes being approximately 35bp. Using the default criteria in MAT, 4784 unique regions-of-interest were detected in at least one of the 32D-C/EBPα-ER clones (n = 3) and absent in control samples 32D-ER (n = 2). Using HAT, the same significance level and maximum fragment size (1 × 10-5 and 600bp respectively) were chosen to detect statistically significant regions-of-interest. Applying these parameters, 1679 statistically significant regions-of-interest were detected in any of the 32D-C/EBPα-ER clones; 80% (1318) of these regions were detected in two or more clones (common regions-of-interest). This corresponds to 856 unique chromosomal regions-of-interest. HAT detected approximately one fifth of the regions-of-interest in comparison with MAT for the same significance level, and 99.9% (855) of these unique detected regions in HAT overlapped with the regions detected by MAT (Figure 2).

To investigate the validity of these detected regions-of-interest (for both HAT and MAT) on the sequence level, a motif enrichment analysis was performed. This was carried out using the Cis-regulatory Element Annotation System (CEAS[79]), where a *P-value* is computed for each known motif, and the motifs that are significantly enriched in the regions-of-interest are reported. The top 10 enriched motifs are indicated in Table 1 for both methods. These data showed that HAT detects regions that are highly enriched for the CEBP motif binding sites, whereas MAT does not show a clear enrichment for these sites. Note that the detected regions-of-interest by HAT, are a subset of MAT.

To investigate detected regions-of-interest based on their flanking genes, regions-of-interest were mapped to the closest 5' transcriptional-start-site of a gene. Mapping is applied on the forward and reverse DNA strands, with a

maximum distance of 300 kb upstream and downstream (NCBI murine genome build 36). This resulted in 2174 unique genes for the 856 unique detected regions-of-interest using HAT (10.7% out of the total set of unique genes present in mouse). These mouse genes were subsequently overlaid with 169 known homologous human C/EBPα target genes (derived from Ingenuity Pathway Analysis, IPA), demonstrating that 40 C/EBPα target genes being detected by HAT ($p \leq 4 \times 10\text{-}7$) and 86 by MAT ($p \leq 3 \times 10\text{-}5$). Note that MAT has detected approximately five times more regions-of-interest (4784) resulting in 7238 unique genes (35.8% out of the total set of unique genes present in mouse). Some of the detected C/EBPα target genes have previously been described, such as: *myc*, *hp*, *mpo* and *il6ra*[52,80-82]. Enrichment of the *il-6 receptor alpha* (*il6ra*) transcriptional-start-site (Figure 3) was subsequently validated by qPCR.

An alternative comparison can be performed using the number of regions-of-interest, instead of the significance level. For HAT; 856 unique regions-of-interest were detected with a significance level $\alpha = 1 \times 10\text{-}5$. To gain approximately the same number of regions-of-interest using MAT, we would need to set the α level at $1 \times 10\text{-}19$, resulting in 893 regions-of-interest. The regions-of-interest detected by HAT showed 84% (718/856) overlap with MAT whereas the overlap of detected regions of MAT with HAT was 83% (742/893). Both methods show a high enrichment for the CEBP binding motifs. Comparing the detected regions-of-interest with respect to MAT (4827 for $\alpha = 1 \times 10\text{-}5$), we need to set the α level higher than 0.05 in HAT, but this may compromise the reliability of detected regions-of-interest. For this reason, we have set the α level at 0.05 and hereby detected 1910 unique regions-of-interest. These were highly enriched for CEBP binding motifs based on the motif enrichment analysis (Table 2), whereas the detected regions-of-interest by MAT were not highly enriched for CEBP binding motifs (Table 1). The regions-of-interest detected by HAT showed 98% (1879/1910) overlap with MAT whereas the overlap of detected regions of MAT with HAT was 39% (1874/4784).

In addition, the HAT and MAT results were also compared with the detected regions of Starr[77]. Starr implements the CMARRT algorithm[76] and thereby incorporates the correlation structure for the identification of regions-of-interest in tiling-array data. For the detection of regions-of-interest, we have utilized similar parameter settings (fragment size = 600bp, minimum number of probes in a region = 8 and $\alpha = 1 \times 105$) as used in HAT and MAT. Using these parameter settings, Starr detected 1664 regions-of-interest and showed high enrichment for CEBP binding motifs (Additional file 1: Supplemental Table S1). Following this, we have examined the overlap of regions-of-interest detected by all methods as depicted in Figure 2. All regions-of-interest detected by HAT (except one) were also detected by MAT alone or together with Starr (64 and 791 respectively). Note that the number of overlapping regions can contain multiple regions-of-interest detected by a single method. To assess the validity of the detected regions-of-interest by HAT, Starr and MAT, we have examined the enrichment for CEBP binding motifs for the different parts in the Venn diagram, depicted as different colors in Figure 2 (blue, red, green, orange and pink). High enrichment for CEBP motifs are found for; *i)* the overlap of HAT with the other two methods (pink: 719), *ii)* the overlap of HAT with MAT (blue: 64) and, *iii)* the overlap between Starr and MAT (orange: 652). No significant enriched motifs are found in the regions detected only by Starr (red: 70) and limited motifs are enriched for CEBP in the regions detected only by MAT (green:

3092). Therefore we can conclude that HAT had the highest specificity as it was able to detect regions-of-interest highly enriched for CEBP binding motifs.

Detection of retroviral insertion sites by HAT

Retroviral Integration Mutagenesis (RIM) in mice is a powerful tool to identify new genes playing an important role in oncogenesis. Mice are injected with retroviruses that potentially integrate into the murine genome upon infection. Viral integration can lead to gene deregulation, and depending on the genes affected, tumors may develop. Genes located proximal to viral integration sites are potentially oncogenic, leading to tumor development. Genomic regions that have been targeted by proviral DNA in multiple tumors are called common viral integration sites (VIS), and are likely driving tumor development. Using retroviral insertional mutagenesis, many oncogenes have been identified using large sequencing screens in multiple tumors[83-86]. We hypothesize that within tumors, genes may be silenced as a result of proviral integration caused by hypermethylation of the CpGs in the viral long terminal repeat, and subsequently in the promoters of their target genes. The identification of methylated genes by means of retroviral insertional mutagenesis may be studied by Methyl-DNA immunopreciptitation (MeDIP-on-chip), followed by inverse PCR, using long terminal repeat (LTR) specific primers. After combining these two technologies, we hybridized samples to Affymetrix promoter chips to identify genomic locations involved in viral integration that potentially harbor new tumor suppressor genes (TSG). Regions-of-interest within this dataset differ from the C/EBPα-study as they have; *i)* a higher variability in fragment sizes and, *ii)* contain specific sequences within the identified regions. Therefore these data are used to examine the performance and broad applicability of HAT among different applications of tiling-array data. Using HAT, we have identified candidate TSGs in mouse tumors by considering regions with a maximum fragment size of 1000bp and a significance level $\alpha = 0.05$. With these parameters, we detected 15 methylated Viral Integration Sites (mVIS); of which one appeared to be a common methylated VIS (cmVIS) among two samples (Figure 4).

Besides the detection of candidate regions based on a statistical framework, we have attached additional mouse genomic sequence information (MM8) to the model, in order to determine the sequence of interest based on the restriction enzyme used in the inverse PCR. Within this assay, a restriction enzyme (DpnII) will cleave DNA at sequence 'GATC', within the integrated viral sequence and the flanking genome. Note that because of this property, it is expected that every detected region must contain a nearby restriction site, which can easily be verified with HAT. HAT showed that all detected mVISs contain a nearby restriction site, conforming specificity of the identified region as being a viral insertion site. For PCR validation of the method, two mVISs were selected based on their location to a nearby 5' transcriptional-start-site, and confirmed. One of the validated regions is illustrated in Figure 5.

Extended applications of HAT

The scope of this method is not limited to the presented studies (i.e., detecting transcription factor binding sites and DNA methylated regions). Moreover, we have successfully applied HAT for the detection of regions enriched for histone modifications such as, trimethylation of histone 3 at lysine 4 or lysine 27 (H3K4 me3 and H3K27 me3) (data

not shown). Some of the detected regions-of-interest were selected for further validation and confirmed by qPCR. Regarding tiling-array data spanning the entire genome[87] (e.g., RNA transcript mapping data[67]), we do not expect changes in algorithm performance (detection of regions-of-interest) due to an increased variability in hybridization consistency because the applied normalization method[74,88] corrects for two major causes of differences in hybridization consistency, i.e., probe sequence and presence of repeats within the genome. Furthermore, in addition to one-color arrays (e.g., Affymetrix tiling-arrays) we envision that HAT can also be applied on data stemming from two-color arrays (e.g., Nimblegen tiling-arrays), because data structure remains similar. We stress however that the normalization procedure is an important step and strongly depends on the type of tiling-array dataset.

CONCLUSIONS

Here we propose a statistical framework; HAT (Hypergeometric Analysis of Tiling-arrays) to analyze tiling-array data. We showed that the method is robust and has increased specificity in the detection of regions-of-interest in comparison with two alternative methods. This is achieved by computing the Hypergeometric probability for every detected region-of-interest, among different threshold levels of probe intensities and window sizes, while keeping control of the Family Wise Error (FWE) by employing Bonferroni correction. Besides the detection of regions-of-interest, HAT also determines sequences-of-interest, flanking genes and the distances to 5' transcriptional-start-sites on both DNA strands. We describe the performance of HAT, when applied to different experimental tiling-array datasets. For each experimental dataset, the selected downstream genes flanking the detected regions-of-interest were successfully confirmed by (q)PCR. We compared the detected regions-of-interest of HAT with two other methods (MAT[74] and Starr[77]), and showed that HAT resulted in a reduced number of detected regions-of-interest using the same significance for both MAT and Starr. However, using motif enrichment analysis we showed that the regions-of-interest detected by HAT were more enriched for the expected binding motifs of CEBP compared to MAT and showed similar enrichment for Starr, illustrating increased specificity using HAT.

Besides analyzing ChIP-on-chip data, HAT is also suitable for the analysis of other types of tiling-array data. Applying HAT to the data from the MeDIP inverse PCR and promoter-GeneChip hybridization experiment, we discovered mVISs and cmVIS that are subject to DNA-methylation and identified the genes (unpublished data) that flank these methylated viral integration sites (Figure 4 and 5).

HAT is applicable to detect regions-of-interest among the different applications of tiling-arrays, and has the advantage of being independent for thresholds, number of probes in a region and probe-resolution. It does not depend on setting various user defined parameters, except for the significance level and an optional maximum fragment size.

METHODS

Extracting candidate gene-regions based on high throughput data using tiling-arrays is a multi-step process (Figure 1). The first step is to normalize the probe intensity data from the chip (Figure 1A). For this purpose, we utilize the normalization from Model-based analysis of tiling-arrays for ChIP-on-chip (MAT)[74,88], but other normalization procedures can also be applied. The normalization procedure prevents systematic variation between experimental conditions, which are unrelated to biological differences. As a result of this normalization, the probe intensity values follow a normal distribution with a negative mean; hence the majority of probes have values below zero, and are ignored in all subsequent analyses. Probe intensities that may be the result of hybridization of labeled DNA on the chip (e.g., were present in the immunoprecipitated chromatin sample), have values greater than zero and are used to determine candidate regions-of-interest.

After normalization, probe intensities are discretized using a varying threshold and the significance of the probes within a varying window is determined. Significant window positions are then merged into the final regions-of-interest. We illustrate this approach in the simplified schematic representation shown in Figure 6. In Figure 6A, eight probes are shown at an arbitrary genomic location. Their intensities are represented by vertical lollipops. The positive probes (six in this example) are assumed to be part of a possible candidate region. Probes with higher intensity levels are more likely to be the results of hybridization on chip, but the exact level of intensity for which this is the case is unknown. Therefore, multiple probe intensity levels are taken into account by varying the discretization threshold t. The number of probes that exceed this threshold (called positive probes) is denoted by $k(t)$. Figure 6B and 6E, illustrates the thresholds $k(t)=2$ and $k(t)=4$, respectively. All probes exceeding t are set to one, and those not exceeding the threshold t are set to zero.

To define a region-of-interest, we determine the significance of all possible window positions g, for which the window contains at least one positive probe. To account for the fact that the exact number of probes in a region-of-interest is undefined, and may differ greatly between different regions-of-interest due to differences in local probe-resolution; the window width n is varied. To prevent evaluating many highly similar windows, thereby incurring a high multiple testing penalty, only those window widths for which the number of probes in the window varies are evaluated. Therefore, n is defined in terms of the number of probes contained in the window. The number of positive probes in a window of width n, at genomic position g, for threshold value t, is denoted by $x(g, t, n)$. In the example presented in Figure 6, we varied n from 1 through to 3. For the case $k(t) = 2$ (Panel B and C), $x(g, t, n)$ ranges from 1 through to 2, and in case $k(t)=4$ (Panel E and F), $x(g, t, n)$ ranges from 1 through to 4. For each window, a *P-value* is determined; defined as the probability of observing at least x positive probes in the window. For any window position g, threshold level t and window width n, $P(g, t, n)$ is computed as:

$$P(g,t,n) = P(X \geq x | g,t,n, X \geq 1) = \frac{P(X \geq x | g,t,n)}{P(X \geq 1 | g,t,n)}$$

(1)

Note; that since we restrict each window to contain at least one positive probe to prevent evaluating useless window positions, this probability is conditioned on $X \geq 1$. All probabilities are computed using the Hypergeometric distribution:

$$P(X \geq x | g, t, n) = 1 - \sum_{0}^{x-1} \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

(2)

Where N is a fixed parameter and represents the total number of probes present on the (e.g., promoter) chip. To correct for the number of tests performed, we apply Bonferroni correction, controlling the Family Wise Error per value of the threshold level as follows:

$$P^*(g, t, n) = P(g, t, n) \cdot k(t) \cdot n$$

(3)

Based on this P-value, it is possible to exclude regions that do not reach a predefined significance level (α):

$$S(g, t, n) = \begin{cases} 1 & \text{if} \quad P*(g, t, n) \leq \alpha \\ 0 & \text{else} \end{cases}$$

(4)

Due to the use of various values for t and n, similar or partly overlapping regions are found. In order to find a single region-of-interest at the same genomic location, these overlapping regions are merged by joining regions with one or more overlapping probes. In our example, we assume for simplicity, that windows with $x\ (g, t, n) \geq 2$ are statistically significant. These statistically significant regions are colored blue and green in Figure 6C and Figure 6F respectively. The merging procedure is illustrated in Figure 6D, where four blue regions are merged into a single region, and in Figure 6G where 18 green regions are merged.

Finally, regions found for different threshold levels t are also merged (Figure 6H) into the final region-of-interest (Figure 6I). Regions-of-interest tend to be larger than the regions detected at a single setting of the threshold level, or single window width due to the merging of all these individual regions. To determine the most important parts of the region-of-interest, we introduce a probe-significance score $Q(g)$, which reports how often probes were part of the statistically significant region. This score is illustrated by the red curve in Figure 6I, and computed as follows:

$$Q(g) = \sum\nolimits_{\forall t} \sum\nolimits_{\forall N} S(g, t, n) \cdot I(x(g, t, n), t)$$

where

$$I\big(x(g,t,n)\big) = \begin{cases} 1 & \text{if} \quad x(g) \geq t \\ 0 & \text{else} \end{cases}$$

(5)

In our example so far, regions are detected within a single sample. When multiple samples are available (for the same experiment), array-wise detection of regions-of-interest is examined in order to detect common regions-of-interest (Figure 1D). A radius, defined in base pairs, can be defined to set the maximum distance between regions over multiple samples (default is zero).

ADDITIONAL PROPERTIES OF HAT

The HAT method includes two additional properties beside the detection of regions-of-interest; *i)* the determination of sequences-of-interest surrounding and within the detected regions-of-interest, e.g., the enhancer binding protein C/EBPα is known to interact with 'CCAAT' sequences, and it is therefore expected that detected regions-of-interest contain this sequence in a chromatin IP experiment. The presence, and positions of the sequences-of-interest can be indicated in the (graphical) output of HAT. In this graphical output, sequences are indicated with an upward facing green bar, indicating that the sequence is detected on the positive strand, or a downward facing green bar representing a sequence on the negative strand. *ii)* The determination of genes flanking the detected regions-of-interest. For every detected region-of-interest (for both upstream and downstream and forward and reverse DNA strands), the genes with the closest distance to the transcriptional-start-site are determined, and indicated in the (graphical) output.

To include these regions-of-interest and genes into the HAT method, the public genome-sequence (available for different model systems) can be utilized from the UCSC genome browser.

AVAILABILITY AND REQUIREMENTS

HAT is implemented in Matlab R2009b and is tested on UNIX and MS-Windows. It is available on http://www.erasmusmc.nl/hematologie/. The run time depends on the number of used threshold cut-offs as the computation complexity increases linear with the used number of probes for the detection of regions-of-interest. In addition, run time also depends on the different steps in the method (Figure 1B-F). On average, for the C/EBPα-study, 28 minutes were needed per sample for the detection of regions-of-interest, while MAT required on average a run time of 23 minutes per sample. Note, however, that in our algorithm the data were analyzed using a multitude of window sizes and thresholds. A more detailed overview of the run time for each step in the method can be found in Additional file 2: Supplemental Figure S1.

AUTHORS' CONTRIBUTIONS

ET, JdR, and MJTR contributed to the conceptual design of the study. ET performed the analysis. RB and BJW performed the biological experiments whereas RB and RD provided biological insights. MJTR, RD, RB, JKP and IPT participated in the discussion of the results. ET, JdR, JKP and RB wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors thank, Erik van den Akker, Martin van Vliet and Mathijs Sanders for the discussions. This research is supported by the Center for Translational Molecular Medicine (CTMM), the Netherlands Genomics Initiative (NGI) and the Dutch Cancer Society (KWF Kankerbestrijding).

FIGURE LEGENDS

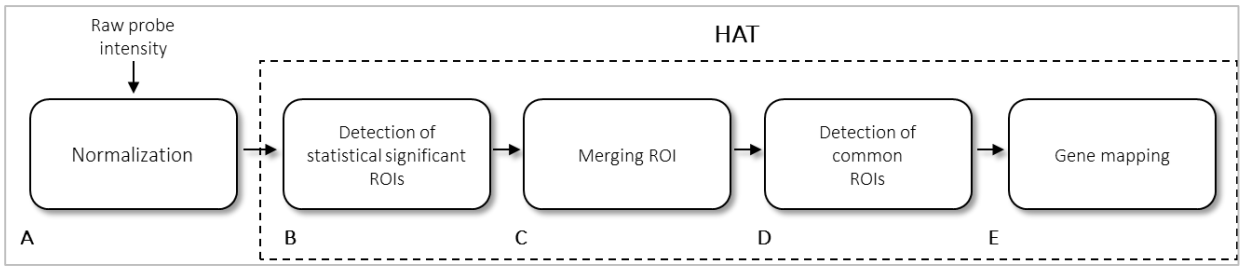


Figure 1. Illustration of the method. The different steps of the method, illustrated as blocks (A, B, C, D and E), are needed to process raw probe intensity data, detection of unique candidate regions and mapping of the detected regions-of-interest to the 5' transcriptional-start-site of nearby located genes. HAT is indicated with the blocks B, C, D and E. These are representative for the detection of unique candidate regions-of-interest in single, as well as multiple samples.

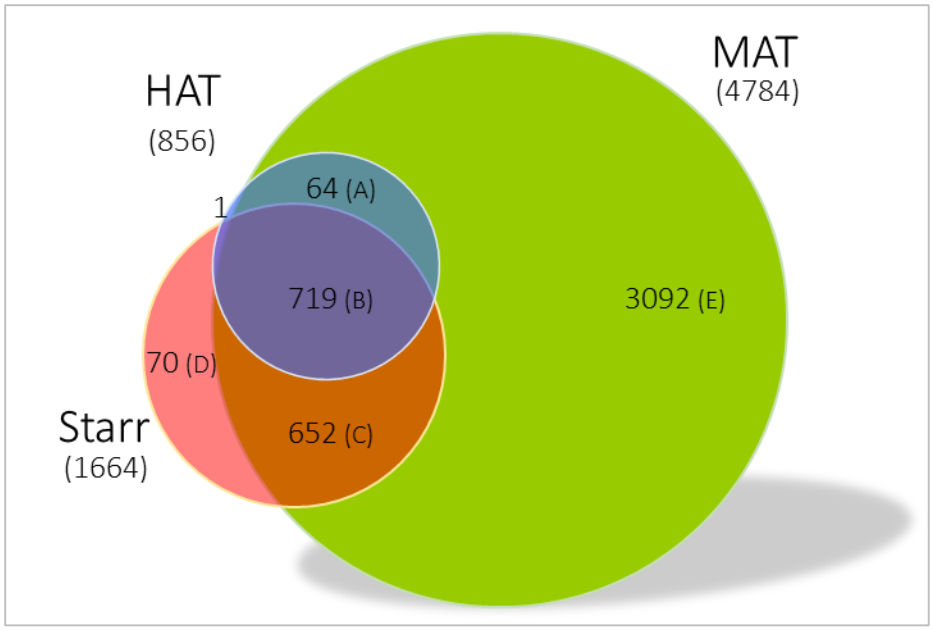


Figure 2. Venn diagram depiction the overlapping regions-of-interest between HAT, Starr and MAT. Detected regions-of-interest by HAT (blue: 856), Starr (red: 1664) and MAT (green: 4784) are indicated with the number of overlapping regions between the methods. The overlap of regions detected by all three methods (pink: 719) showed high enrichment for CEBP binding motifs. Overlapping regions between HAT and MAT (64: blue) and Starr and MAT (orange: 652) also showed high enrichment for CEBP binding motifs. Uniquely detected regions by Starr (red: 70) showed no significantly enriched motifs, and MAT (green: 3092) showed limited motifs enriched for CEBP. Note that the number of overlapping regions can contain multiple regions-of-interest detected by a single method.

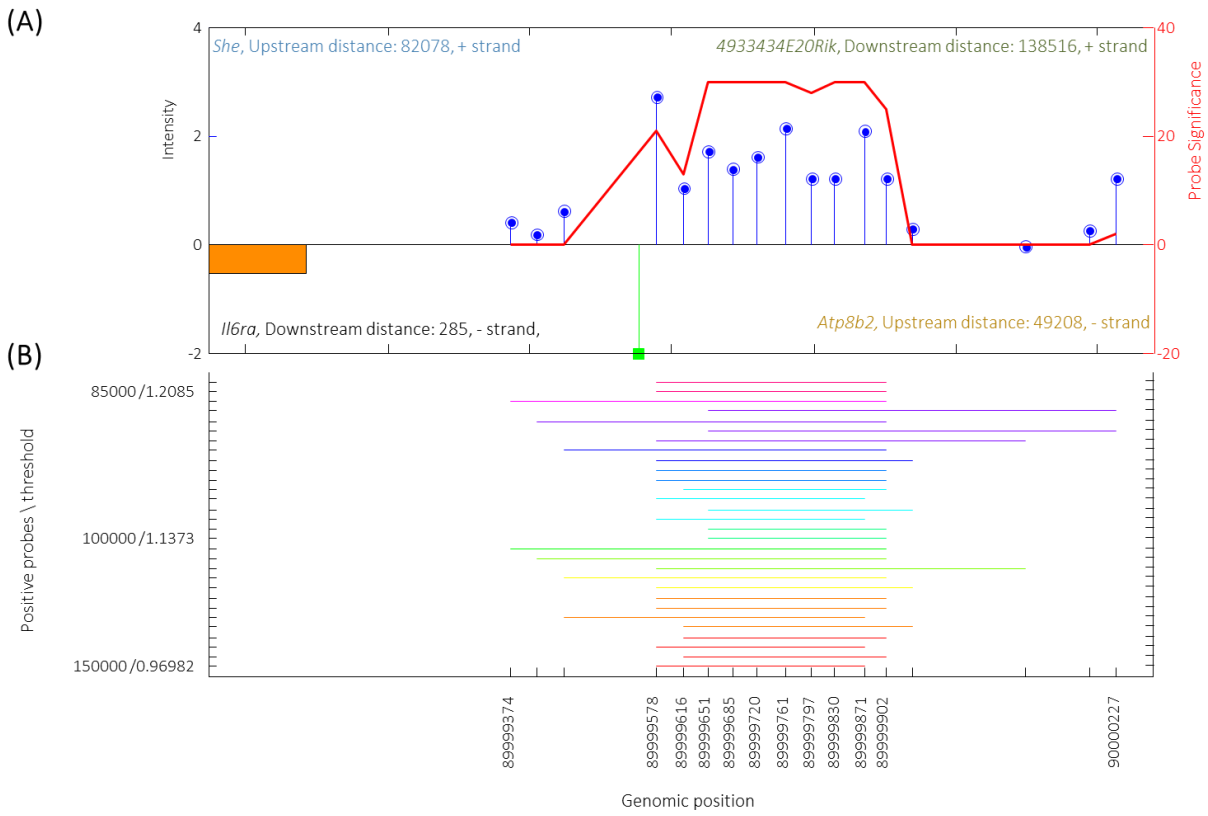


Figure 3. Graphical output of a detected region-of-interest from the C/EBPα-study. It was confirmed with qPCR that the C/EBPα protein targets and regulates the proximal promoter region of the *il-6 receptor alpha* gene, which lies downstream of the region-of-interest (negative DNA strand). The top panel (A), indicates the probes, represented as vertical blue lollipops, the left y-axis the probe intensities, and the right y-axis illustrates the contribution of each probe separately to the region (probe-significance). The x-axis indicates the genomic probe positions, and illustrates with a downwards facing green bar; the sequence of interest. The sequence, 'CCAAT', was found on the negative DNA strand. Furthermore, flanking genes to this detected region are indicated with distances in base pairs to the 5' transcriptional-start-site. In the bottom panel (B), the detected regions-of-interest for various windows and probes are shown. The colors represent the detection of regions-of-interest, for a number of different top probes and window sizes. The merged region-of-interest has a fragment width of 853bp, and lies in the proximal promoter region of *il6ra* on the negative DNA strand.

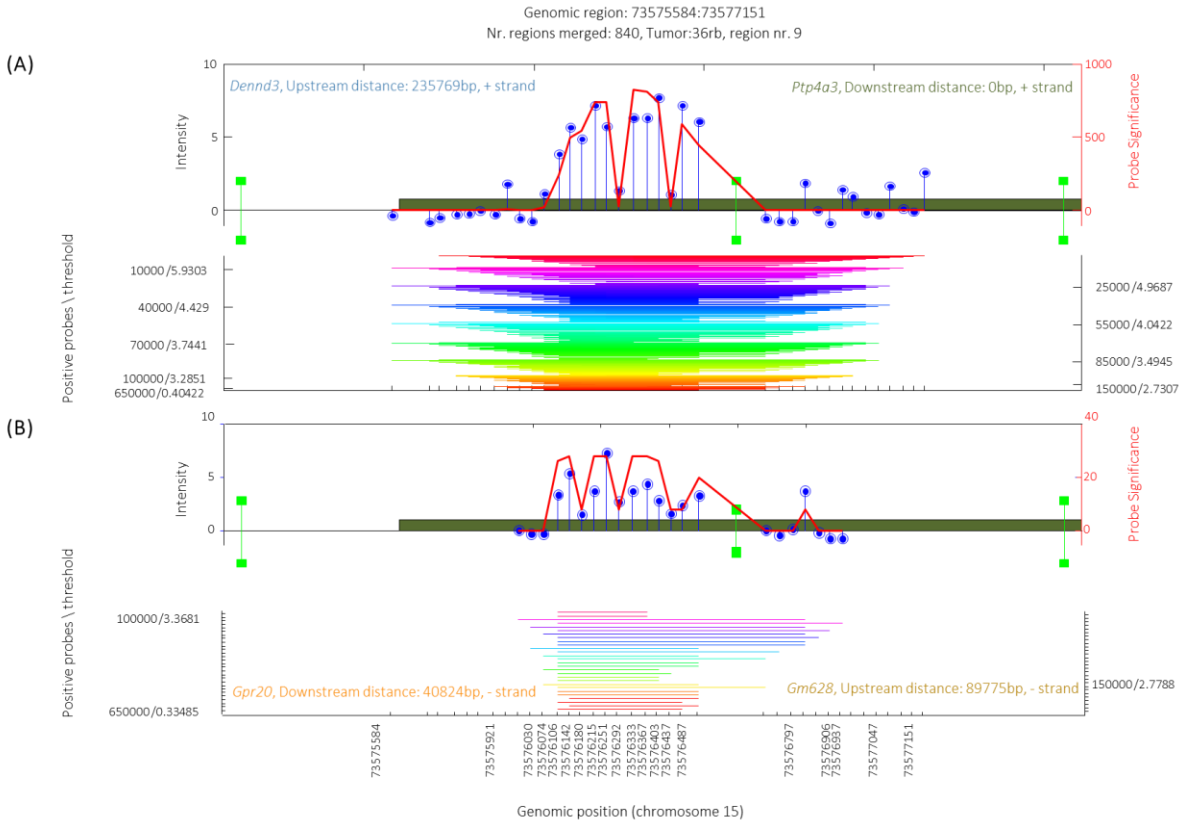


Figure 4. Graphical output of a detected cmVIS in the MeDIP-study. A region-of-interest detected in two samples, is illustrated in Panels A and B. Panel A shows 840 subregions that are merged with a total length of 1567bp. The restriction sites, indicated as green bars, are located in and around the detected region, and are present on both DNA strands due to the palindrome sequence: 'GATC'. The region-of-interest detected in the second tumor (Panel B), exists of 28 subregions, with a fragment width of 949bp.

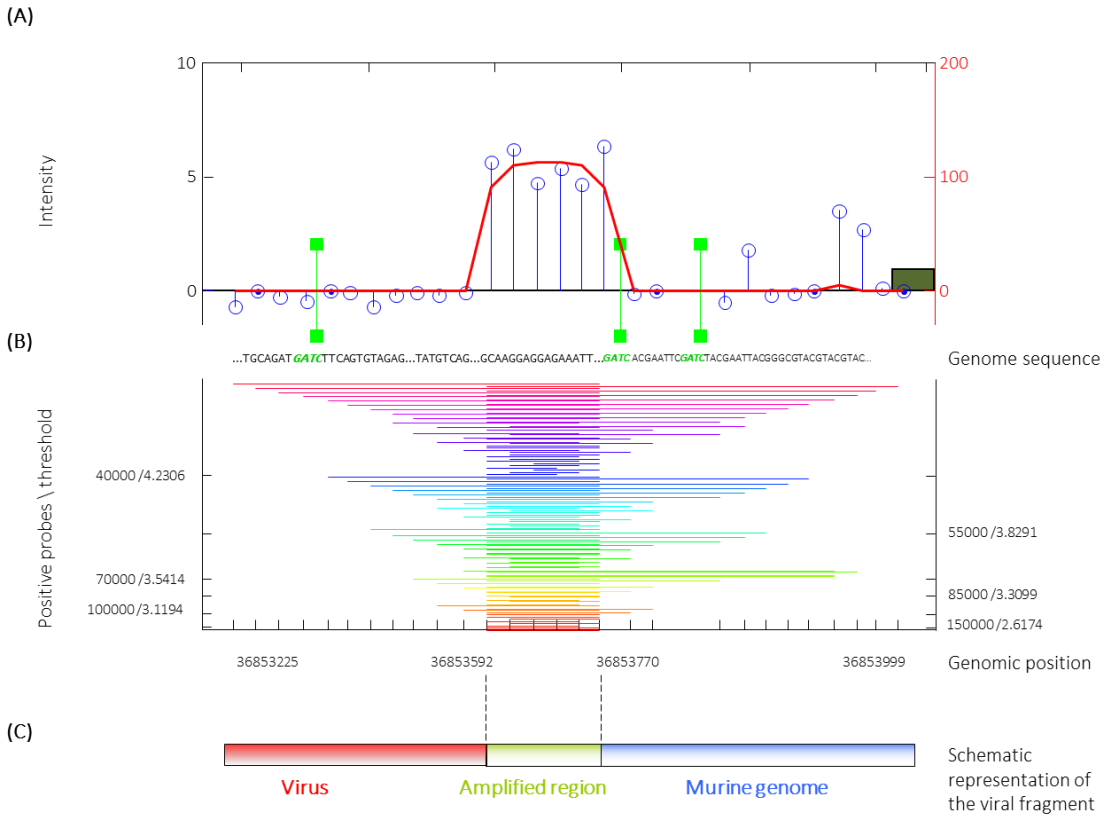


Figure 5. Graphical output of a detected and validated mVIS in the MeDIP-study. Panel A illustrates the detected mVIS which are subject to DNA-methylation. Only a section of the detected region-of-interest has an increased probe intensity; the probe-significance signifying this subregion. Directly beside the increased probe-significance, a restriction cleavage site is indicated by means of a green bar. Due to the palindrome sequence, these sites are indicated at the same genomic position on both DNA strands. Panel B shows the detected statistically significant regions among the different thresholds, and window sizes with various colors. A schematic representation of the amplified genomic region, with the virus- and the murine contribution, is shown in Panel C.

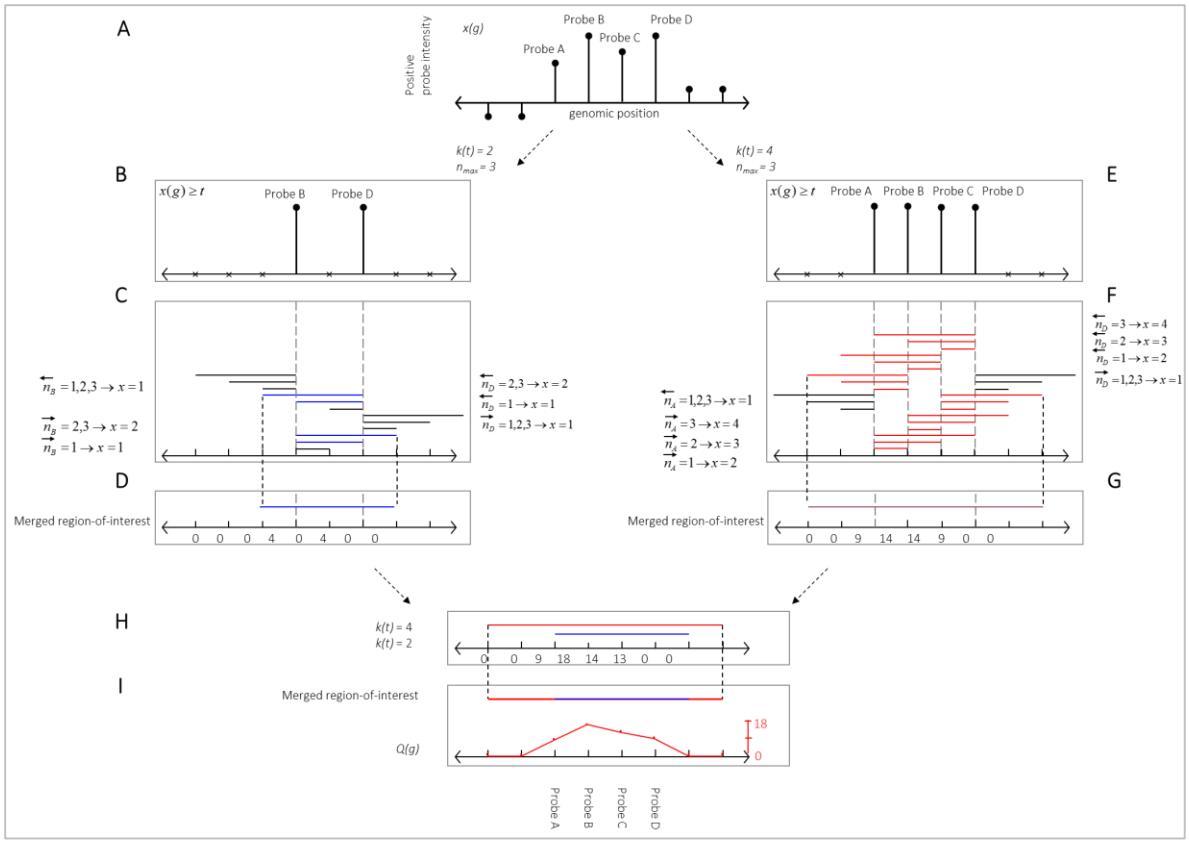


Figure 6. Schematic depiction for the detection of regions-of-interest. Schematic depiction for the detection of regions-of-interest, based on probe intensities. Eight probes, with their genomic location, are shown in Panel A. Four of these have positive probe intensities. The use of multiple thresholds, transforms continuous data into discrete data; as shown in Panel B and E. Various window scales *N*, are used to examine neighboring probes for their probe intensities in Panel C and F. These windows will contain different number of positive probes. The Hypergeometric probability is computed for every region-of-interest, and excludes a region-of-interest when the region is not statistically significant after correcting for a single positive probe in a region-of-interest and multiple testing. The remaining regions are merged for each *k(t)* (illustrated in Panel D, G, H) and then among all *k(t)* to a single region-of-interest (Panel I). To determine how often probes were detected in statistically significant regions, the probe-significance is computed (Panel D and E), and indicated with a red colored line that signifies the statistically significant probes in the detected region-of-interest.

	MAT				HAT			
Nr	Motif	Hits	Fold-change	*p* -value	Motif	Hits	Fold-change	*p* -value
1	AP2alpha	9735	1.606	0	M00117.CEBPbeta	1532	2.325	2.84E-185
2	Elk-1	5380	1.707	9.23E-286	M00770.CEBP	3076	1.766	2.23E-183
3	M00470.AP-2 gamma	5938	1.641	1.82E-274	M00912.C-EBP	3036	1.715	1.31E-164
4	M00109.CEBPbeta	6170	1.617	3.52E-269	cEBP	1928	1.965	1.89E-157
5	M00695.ETF	3449	1.885	3.05E-250	M00116.CEBPalpha	2689	1.722	4.30E-148
6	M00025.Elk-1	2979	1.949	1.76E-237	M00109.CEBPbeta	1278	2.161	2.35E-132
7	M00446.Spz1	4863	1.665	3.90E-237	M00190.CEBP	2402	1.719	9.67E-132
8	M00008.Sp1	5135	1.625	1.04E-228	M00098.Pax-2	1799	1.578	8.57E-73
9	E74A	3635	1.691	7.08E-188	M00496.STAT1	1909	1.545	8.88E-71
10	M00771.ETS	3756	1.674	2.37E-187	M00971.Ets	1917	1.508	4.42E-64

Table 1. Motif enrichment analysis. The top 10 motifs enriched in the detected regions-of-interest ($\alpha = 1 \times 10^{-5}$) by HAT and MAT for the C/EBPα-study (ChIP-on-chip). Among the top 10 motifs enriched in the regions-of-interest detected with HAT, seven contained the CEBP binding motif whereas for MAT, only one contained the CEBP binding motif. For each reported motif, the number of hits within the regions-of-interest are counted, their fold change computed, and the *P-value* derived using the binomial test.

Nr	Motif	Hits	Fold-change	*p* -value
1	M00117.CEBPbeta	3236	2.082	6.19E-304
2	M00770.CEBP	6688	1.628	8.95E-299
3	M00912.C-EBP	6609	1.583	6.70E-265
4	M00116.CEBPalpha	5858	1.591	1.12E-239
5	cEBP	4068	1.758	1.88E-238
6	M00190.CEBP	5245	1.592	2.81E-215
7	M00716.ZF5	3927	1.706	2.12E-208
8	M00109.CEBPbeta	2645	1.896	3.06E-195
9	M00098.Pax-2	4355	1.619	1.76E-191
10	M00428.E2F-1	4374	1.572	4.67E-171

Table 2. HAT: Motif enrichment analysis using α = 0.05. The top 10 motifs enriched in the 1910 detected regions-of-interest using HAT ($\alpha = 0.05$) in the C/EBPα-study. There is a high enrichment for binding motif CEBP. For each reported motif, the number of hits within the regions-of-interest are counted, their fold change computed, and the *P-value* derived using the binomial test.

SUPPORTING MATERIAL

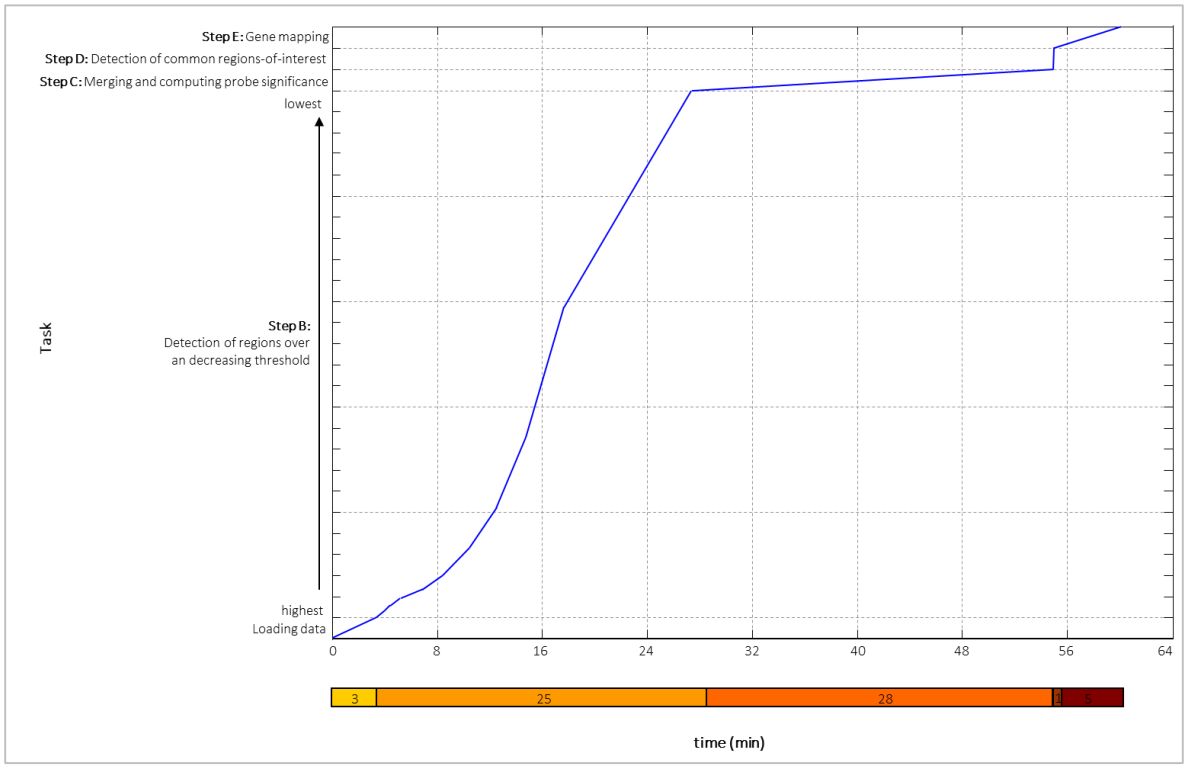


Figure S1. HAT Computation performance. Run time of the various steps in the method. The C/EBPα-study is used to analyze the run time for the different steps in the method; Step B: loading data and detection of regions-of-interest, Step C: Merging of regions-of-interest and computation of the probe-significance, Step D: detection of common regions-of-interest and Step E: gene mapping. Per sample, 62 minutes were needed on average to process all the steps in the method.

Nr	Motif	Hits	Fold-change	*p*-value
1	M00117.CEBPbeta	1992	2.153	2.74E-203
2	M00912.C-EBP	4084	1.643	9.46E-190
3	cEBP	2570	1.866	2.94E-181
4	M00770.CEBP	3980	1.627	1.90E-178
5	M00116.CEBPalpha	3523	1.607	7.15E-151
6	M00109.CEBPbeta	1670	2.011	5.45E-145
7	M00190.CEBP	3138	1.599	3.13E-132
8	HLF	666	2.023	3.80E-60
9	M00260.HLF	608	2.017	9.94E-55
10	M00771.ETS	845	1.731	6.80E-49

Table S1. Starr: Motif enrichment analysis. The top 10 motifs enriched in the 1664 detected regions-of-interest using Starr (fragment size = 600bp, minimum number of probes in a region=8, α=1×10^{-5}) in the C/EBPα-study. There is a high enrichment for binding motif CEBP. For each reported motif, the number of hits within the regions-of-interest are counted, their fold change computed, and the *P-value* derived using the binomial test.

CHAPTER 3

HATSEQ: Detection and Interpretation of Peaks in Tiling-array and Sequence Data

Advances and Applications in Bioinformatics and Chemistry

In press

HATSEQ: Detection and Interpretation of Peaks in Tiling-array and Sequence Data

Erdogan Taskesen, Remco Hoogeboezem, Ruud Delwel and Marcel J.T. Reinders

ABSTRACT

Probing protein-DNA is gaining popularity as it sheds light on molecular mechanisms that regulate the expression of genes. Currently, tiling-arrays and next-generation sequencing technology can be used to measure these interactions. Both methods generate a signal over the genome in which contiguous regions of peaks on the genome represent the presence of an interacting molecule. Many methods do exist to identify functional regions-of-interest (ROIs) on the genome. However the detection of ROIs are often not an end-point in research questions and it therefore requires data dragging between tools to relate the ROIs to information present in databases, such as gene-ontology, pathway information, or enrichment of certain genomic content.

We introduce HATSEQ (Hypergeometric Analysis of Tiling-array and Sequence data), a powerful tool that accurately identifies functional ROIs on the genome where a genomic signal significantly deviates from the general genome wide behavior. HATSEQ also includes a number of built-in post-analyses with which biological meaning can be attached to the detected ROIs in terms of gene pathways and dé-novo motif analysis, and provides different visualizations and statistical summaries for the detected ROIs. On top of that, HATSEQ has an intuitive graphic user interface that lowers the barrier for researchers to analyze their data without the need of scripting languages. We compared the results of HATSEQ against two other popular ChIP-Seq methods and observed overlap in the detected ROIs but HATSEQ is more specific in delineating the peak boundaries. We also discuss the versatility of HATSEQ by using a STAT1 ChIP-Seq data set, and show that the detected ROIs are highly specific for the expected *STAT1* binding motif. HATSEQ is freely available at: http://hema13.erasmusmc.nl/index.php/HATSEQ.

BACKGROUND

Protein-DNA-interactions, such as transcription factor-DNA-binding, DNA-methylation or methylation/acetylation of histone tails, can nowadays be identified with high sensitivity and specificity, using next-generation sequencing (NGS) technology. NGS rapidly replaces tiling-arrays technology because of the increased resolution with which the interactions can be measured. Both technologies generate a signal along the genome that for instance represents the interaction of regions with transcription factors. Typically one is interested in finding those regions in the genome where a signal significantly deviates from the overall genome wide background signals. Previously, for tiling-array data, we developed a method called “Hypergeometric Analysis of tiling-arrays” (HAT), to detect regions-of-interest (ROIs). In short, HAT sets a threshold to decide whether the signal of a probe is excessive, and then uses a sliding

window approach to analyze whether a significant number of marked probes are found within that window. The signal is analyzed at different scales by considering a range of different thresholds and window sizes, and the detected regions at individual scales are integrated. The detected ROIs are over all scales under control of a Familywise error (FWE), specified by a significance level α. HAT has been successfully applied on a range of different DNA-interaction sources, such as ChIP-on-chip[89], MeDIP-on-chip[56], H3K4me3, H3K27me3[89] and 3′-TILLING-135-K-*Oryza-sativa*-microarray.[57] Here, we introduce HATSEQ, which is an improved version of HAT that can work on nucleotide resolution. As with HAT, HATSEQ is nonparametric, and independent of the coverage and resolution across the genome. Various methods with varying algorithmic complexity have been developed to detect ROIs in ChIP-Seq data such as, MACS[90], FindPeaks[91], CisGenome[92], QuEST[93], PeakSeq.[94] MACS (Model-based Analysis for ChIP-Seq) is one of the most cited methods for analyzing ChIP-Seq data. Although the variety of ChIP-Seq methods, the majority can only be run from the command line and require variable degrees of data formatting and expertise to implement.[95] CisGenome however does provides a graphical user interface (GUI) but is restricted to the windows platform. With HATSEQ, we aim at the researcher that can experience difficulties with the use of the command line and in the downstream analysis. After finding the ROIs with HATSEQ, one is generally interested in a functional analysis of the regions. Typically this is done by relating the regions to information present in databases, such as gene-ontology, pathway information, or enrichment of certain genomic content. HATSEQ supports, through a GUI, a number of such functional analyses of the ROIs: e.g., gene mapping, motif analysis and pathway analysis. It also outputs for the detected ROIs, fasta files, UCSC genome browser tracks to enable visualization of the ROIs together with any other genomic data, and a single circular graph (Circos[96]) that illustrates all the detected genes and their chromosomal locations.

IMPLEMENTATION

HASTSEQ: a statistical framework to detect regions-of-interest in genomic signals

HATSEQ detects ROIs in NGS data using the statistical framework as described in HAT[89], but with read depth at genomic positions as an input. It is supposed that genomic positions with read depth greater than zero may be the result of sequenced DNA pieces that were, e.g., present in the immunoprecipitated chromatin sample, indicating the presence of protein-DNA-binding at that particular position. To decide whether the read depth at a genomic location is excessive, HATSEQ varies the threshold at which it considers the read depth to be indicative for a genomic event. A sliding window approach is then used to analyze whether a significant number of excessive sequence-reads are found within the window for every threshold setting and for varying widths of the window (as the size of the event is not known a priori). For each window, a P-value is determined, defined as the probability of observing at least the number of observed reads*, x,* in the window (given a random distribution of reads over the genome). For any window position g, threshold level t and window width n, $P(g, t, n)$ is computed as:

$$P(g,t,n) = P(X \geq x|g,t,n,X \geq c) = \frac{P(X \geq x|g,t,n)}{P(X \geq c|g,t,n)}$$

(1)

Where $P(X \geq x \mid g, t, n)$ is based on the Hypergeometric distribution of drawing, on genomic position *g*, at least *x* reads that exceed the threshold *t* in a window of size *n,* and where *N* is a fixed parameter that represents the total number of reads that are sequenced, and *K* the number of reads that exceed the threshold. For each window the P-value is restricted such that each window should contain at least *c* reads to prevent evaluating window positions that are not of interest.

We apply Bonferroni to correct for the number of tests performed at each threshold level, which is defined by the number of reads (*K)* that exceed the threshold (*t)* and window size *n.* The corrected P-values are subsequently defined by: $P^*(g, t, n)$. Due to the use of various threshold values (*t)* and window sizes (*n)*, similar or partly overlapping regions are found. In order to find a single region-of-interest at the same genomic location, these overlapping regions are integrated by joining regions with one or more overlapping reads. To determine the most important part of the region-of-interest, we introduce a read depth significance score $Q(g)$, which reports how often reads were part of a region for a predefined significance level (α). This score is computed as follows:

$$Q(g) = \sum_{\forall t} \sum_{\forall N} S(g,t,n) \cdot I(x(g,t,n),t)$$

where

$$S(g,t,n) = \begin{cases} 1 & \text{if } P^*(g,t,n) \leq \alpha \\ 0 & \text{else} \end{cases} \quad \text{and} \quad I\big(x(g,t,n)\big) = \begin{cases} 1 & \text{if } x(g) \geq t \\ 0 & \text{else} \end{cases}$$

(2)

Thus, the final candidate regions-of-interest are determined by integrating the significant window positions over all thresholds. HATSEQ is optimized for NGS data analysis by: (1) incorporating a minimum allowed read depth to prevent the detection of systematic variation; (2) incorporating a minimum allowed region length to prevent the detection of regions that are the result of highly correlated reads; (3) normalization of the read depth per sample such that sum of the read depth is 1, which makes the depth of the sequenced reads comparable between experiments; (4) normalization of the read depth by using a set of reference samples ; and (5) the use of multi-threaded computations (each chromosome is separately analyzed and HATSEQ exploits the use of memory mapped files that allows the analysis of any read depth).

HATSEQ can be applied in three types of study-designs, namely; (1) one sample analysis where only one sample is available and sequenced; (2) multi-sample analysis, where the sequenced reads of the experimental samples can be analyzed compared to the reads of one or more negative control samples; and (3) combined ChIP-Seq and ChIP-on-chip analysis where an overlap of candidate ROI between the experimental replicates can be marked.

Functionalities of HATSEQ

Data processing and region identification. HATSEQ detects ROIs from mapped sequenced reads or normalized probe intensities. For the analysis of NGS data, it processes Bam or Pileup files to detect ROIs using the read depth at base pair position. For ChIP-on-chip data it uses preprocessed files, e.g., by MAT.[74] As an example, both NGS data and ChIP-on-chip files can be loaded using the GUI and simultaneously analyzed with or without controls.

Pathway analysis. HATSEQ integrates two pathway enrichment analyses based on the genes that are selected by; (1) having a selected ROI as closest ROI; or (2) having a detected ROI in their promoter region (the 2000bp region upstream of the transcriptional-start-site (TSS)). Pathway annotations (gene sets) are extracted from the Molecular Signature Database[97] (MSigDB). The enrichment of each pathway for the selected set of genes is computed using the Hypergeometric distribution and is corrected for multiple testing using FDR[98] or FWER.[99]

Motif analysis. HATSEQ gives the opportunity to find enriched motifs in sequences derived from: (1) the detected ROIs; and (2) the promoter regions (2000nt upstream from TSS) of the genes that have a selected ROI as closest ROI. It uses the generalized extreme value probability method[100], which detects significantly over-represented ungapped words of fixed length. It consequently outputs the over-represented sequences that are corrected for multiple testing using FDR[98] or FWER.[99] Finally, for each detected motif, the position weigh matrices (PWMs) are correlated with annotated PWMs from TRANSFAC and JASPAR and subsequently listed if the correlation is larger than 0.6.

Support for different species. HATSEQ supports gene-annotation (for e.g., ROI gene-associations) and chromosome files for the species that are available on UCSC (http://hgdownload.cse.ucsc.edu). Species that are available on UCSC can be chosen using the GUI, which are then automatically downloaded, or alternatively, species can be uploaded selectively.

Statistical summaries and visualization of results. HATSEQ reports the detected ROIs, including the neighbouring genes and summary statistics, in tables. For example, one can extract the percentage of ROIs that are in close vicinity to the TSS of a gene, or the percentage of ROIs that contain a user defined motif. The genes for the detected ROIs can be visualized by the circular graph, Circos or as custom tracks in UCSC.

Equipment-Software. HATSEQ is a stand-alone application that is implemented in C++ and Matlab Mathworks. To run HATSEQ, an installation of Matlab or the freely available Matlab Compiler Runtime (MCR) is mandatory.

Equipment-Hardware. HATSEQ runs on any x86-64 system with MS-Windows, UNIX, Linux or Mac OS whereas a minimum of 4GB RAM is required. The analyzed ChIP-Seq examples in this manuscript wererun on MS-Windows 7 with a 1.87-GHz CPU and 4GB RAM. The runtime, with default parameter settings, was approximately 10 minutes to detect ROIs in 1 Million reads (1.87GHz), an estimate that increases with sequence coverage.

RESULTS AND DISCUSSION

Method comparison

To evaluate the performance of HATSEQ, we used two publically available ChIP-Seq data sets (DNA-binding to CCAAT enhancer binding protein alpha (C/EBPα) and trimethylation of H3 lysine 4 (H3K4me) experiment) and compared the results against to two other state-of-the-art methods, i.e., MACS (v1.42)[90] and FindPeaks (v4).[91] MACS uses a dynamic Poisson distribution to detect peaks and empirically estimates the false discovery rate (FDR) for each detected peak, whereas FindPeaks assumes a triangle based distribution in which fragments have a minimum, maximum and a user defined median size.

The first ChIP-Seq data set contains massively parallel sequenced DNA-fragments bound by the transcription factor C/EBPα (cell line U937, GEO accession: GSM722423) and is used to evaluate the results for one sample analysis. The sequencing data of this C/EBPα experiment is aligned using BWA[101] (hg19). To avoid the detection of peaks that are the result of technical variation, we discarded genomic positions with a read depth smaller than 10. With MACS we detected 50,525 ROIs, using default parameters (bandwidth of 300bp at the $1x10^{-5}$ significance level). FindPeaks detected 75,839 ROIs using the default parameters (Triangle distance low=100bp, median=200bp, high=300bp with minimal allowed coverage 0.001). With HATSEQ we detected 32,735 ROIs using a bandwidth (fragment size) of 300bp, but with FWER significance level 0.05. Eighty-seven percent of the 32,735 HATSEQ ROIs (28,413 ROIs) were also detected by either of the two other methods, and 85% (27,862 ROIs) of the HATSEQ ROIs are common among all methods (Figure 1A).

Although there was a high overlap of detected ROIs between the three methods, HATSEQ better delineates the peak boundaries in the data. This can be concluded from: (1) regions detected by HATSEQ showed on average higher read depth (HATSEQ: 30.1, MACS: 13.1 and FindPeaks: 5, Figure 2C); (2) regions detected by HATSEQ are consistently smaller in length compared to the other methods (average region length HATSEQ: 153bp, MACS: 350bp and FindPeaks: 1,679bp, Figure 2A); and (3) the read depth differences at the boundary of a region are more extreme for HATSEQ regions (Figure 2B). We illustrate in Figure 1B the superior behavior of HATSEQ for ChIP-seq data for a region on chromosome 1 of the C/EBPα experiment. It can clearly be seen that HATSEQ most accurately detects the three regions-of-interest, among a region close to the TSS of *IL6R* which is a known target of C/EBPα.[82] Remarkably, FindPeaks detects one large region-of-interest, and MACS overshoots the boundaries of the three regions. Among the 4,322 ROIs that were solely detected by HATSEQ, we detected ROIs that were in close proximity of known target C/EBPα genes, such as *CD7*[61] and *ACSL*[102].

The second analysis involved sequence data from a H3K4me ChIP-Seq experiment (cell line K562, data available from University of Washington[10]) in which functional loci based on the chromatin signatures can be identified, i.e., H3K4me peaks at the promoter of active genes.[103] These histone marks are known to generate a bimodal distribution of the signal (read depth) which is caused by the spacing between the histones that interact with the DNA.[104] We evaluated the results of HATSEQ, MACS and FindPeaks for the identification of H3K4me peaks by normalizing it against a control replicate. Sequence alignment was performed using BWA[101] (hg19) with default parameter settings. HATSEQ

detected 14,616 statistically significant regions-of-interest, MACS: 10,694 and FindPeaks: 9,471 (Figure 1C) by comparing the input versus the negative control.

The regions detected by HATSEQ that overlap with either of the two other methods (9,286 ROIs, 63.5%) again showed that HATSEQ better delineates the peaks, although less pronounced as in the previous experiment: (1) the HATSEQ regions have higher read depths (average read depth: HATSEQ: 16.1, MACS: 15.5 and FindPeaks: 10.5, Figure 3C), (2) HATSEQ regions are smaller in length (average region length: HATSEQ: 1,096bp, MACS: 1,751bp and FindPeaks: 4,297bp, Figure 3A), and (3) the difference of read depth at the border of the region are much more pronounced for HATSEQ regions (Figure 3B). Figure 1D illustrates a region on chromosome 22 in close proximity of *SDF2L1.* Clearly HATSEQ delineates the boundaries of the peak region best. To assess the validity of the detected regions by HATSEQ we tested the 14,616 ROIs for bimodality using the statistical dip test of unimodality.[105] A significant bimodal distribution (FDR≤0.05) was detected in 12,897 ROIs (88.2%). This illustrates that the large majority of detected ROIs contains the expected bimodal distribution.

Taken together, HATSEQ showed better performance in delineating peak boundaries for the detected ROIs when compared to other ChIP-Seq methodologies, such as MACS and FindPeaks. For each method we used the default settings, although transcription factor binding and histone modifications can differ substantially in their properties (e.g., length or the region) yet specifying the optimal parameters in an unbiased way is difficult. We also tested whether HATSEQ can also detect ROIs in genomic areas with low read depth by re-analyzing the C/EBPα ChIP-seq data set without removing any genomic positions with read depth smaller then 10. We detected 42,046 significant regions (instead of 32,735 ROIs) which clearly illustrates the capability of HATSEQ to detect ROIs in low read depth genomic areas. Note that applications of HATSEQ are not limited to the presented NGS ChIP data but can be applied to other types of data, such as MeDIP-seq[106], DNase-seq[107] and MBD-seq.[108]

A case study with HATSEQ

To illustrate the functionalities of HATSEQ, we used a publicly available ChIP-Seq data set (GEO accession: GSE15353) where the DNA-fragments bound by the transcription factor STAT1[109] were massively parallel sequenced. For transcription factor STAT1 it has been described that it binds to STAT-motifs[110], and a well-known target gene is the *STAT3*[111] gene. We compared data obtained from six interferon-γ (IFN-γ) stimulated HeLa S3 cells and compared those to seven unstimulated human HeLa S3 cells. After the alignment using BWA[101], we detected in total 2,502 ROIs with HATSEQ (sizes between 11bp-669bp, median: 81bp) using default parameter settings (α≤0.05 and read depth≥10). These ROIs showed significant binding in the stimulated cells but not in the unstimulated cells, which were subsequently investigated using HATSEQ's motif analysis. Thus, from the design of the experiment, it is expected that the detected ROIs should contain STAT-binding sites. The detected motifs, among the sequences of the 2,502 ROIs correspond to the STAT1 motif according to our results (*P-value*<9.1×10^{-6}), and also according to MEME[112] and TOMTOM.[113] The 2,502 detected ROIs are annotated with 914 unique genes. These 914 genes included the *STAT3*

gene, which was associated with one of the most significantly detected ROI. This ROI was also strongly enriched for the STAT1 motif sequence (P-value<2.13x10-177, Figure 4A and B). However, not all detected ROIs contained the STAT-binding site. Therefore we searched for ROIs that were detected across two or more replicates. We found 511 ROIs that were consistently detected, i.e., in two or more replicates (Figure 4D). The HATSEQ motif analyses on these 511 consistently detected ROIs showed a strong enrichment for the STAT-binding site (Figure 4C), and was seen in 88% of these ROIs. In addition, using HATSEQ we found 47 enriched MSigDB pathways for these 511 ROIs including a pathway that involve *STAT3* and its targets (Figure 4E).

CONCLUSIONS

We present HATSEQ, a tool to analyze both tiling-array and NGS data. We applied HATSEQ to analyze a STAT1 ChIP-Sequence experiment and detected ROIs that were enriched for the STAT1 motif. In addition, we detected unknown as well as previously reported direct target genes of STAT1*: STAT2*[114]*, STAT3*[111]*, IRF1*[115]*, IL-27*[116]*, PTK2*[117] and *IFNAR2.*[118] HATSEQ can be used for single sample analysis or with a set of reference samples whereas the expected regions-of-interest can be of any size. We showed for both the C/EBPα and H3K4me ChIP-Seq experiments that HATSEQ better delineates the peak boundaries. HATSEQ is a powerful tool with an intuitive graphic user interface that lowers the barrier for researchers to detect regions-of-interest in genomic signals, and integrates an analysis of these detected regions to enhance their functional role.

AVAILABILITY

The HATSEQ program is freely available on http://hema13.erasmusmc.nl/index.php/HATSEQ or http://www.erasmusmc.nl/hematologie/. The required Matlab Compiler Runtime (MCR) executable is provided.

AUTHORS' CONTRIBUTIONS

ET designed and developed the software, analyzed and interpreted the data, and drafted the manuscript. ET, RD and MJTR participated in the design of the study and contributed to the writing of the paper. RH implemented an enhanced version of the back-end. All authors provided relevant input at different stages of the project, and read and approved the final manuscript. This research is funded by the Center for Translational Molecular Medicine, project BioCHIP (grant 03O-102).

ACKNOWLEDGEMENT

We thank F. Lesmana for her contribution in the development of the Graphic User Interface (GUI).

FIGURE LEGENDS

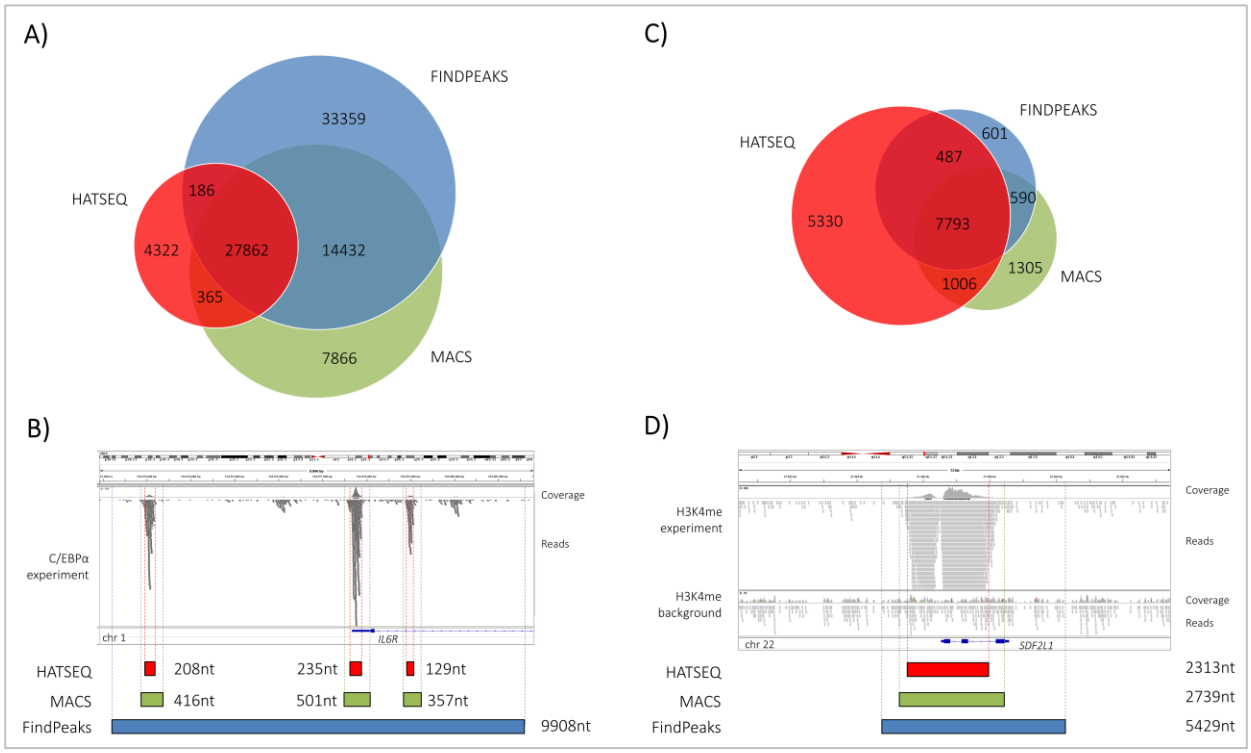


Figure1. Venn diagram and an illustration of a detected ROI for HATSEQ, MACS, and FindPeaks. Detected regions-of-interest by HATSEQ, MACS and FindPeaks are indicated in red, green and blue respectively. (A) The amount of detected ROIs for the C/EBPα experiment, and the overlap between the methods. (B) ROIs detected by the three methods on chromosome 1 (around the promoter region of *IL6R*). The top part of this panel illustrates the pileup or coverage that is determined by the sequenced reads. (C) The amount of detected ROIs for the H3K4me experiment and the overlap between the methods. (D) ROIs detected in the neighbourhood of *SDF2L1* on chromosome 22. The top panel of this Figure shows a pileup of the H3K4me experiment as well as a pileup of an H3K4me background experiment (giving an indication of the amount of non-specific reads).

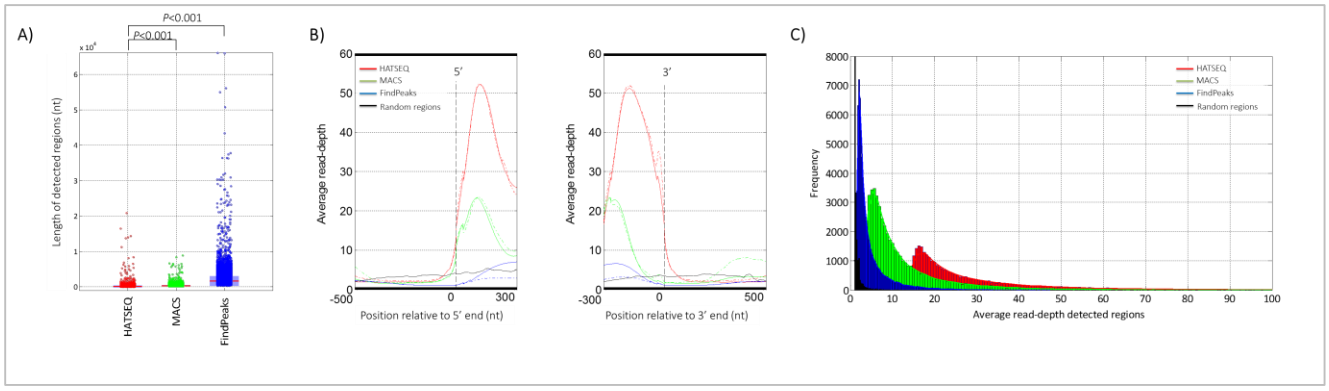


Figure 2. ROI statistics for the C/EBPα experiment. Statistics for the detected ROIs by HATSEQ, MACS and FindPeaks (red, green and blue respectively) for the C/EBPα experiment. (A) Boxplot illustrating the region length of the detected regions. (B) Average read depth across the ROI boundaries with respect to the 5' (left panel) and 3'end (right panel). The average read depths are calculated per nucleotide position after aligning the detected ROIs at their 5' and 3' ends, respectively. The solid line represents the alignment of the 32,735, 50,525 and 75,839 ROIs detected by HATSEQ, MACS and FindPeaks respectively. The dashed line represents the alignment of the 4,322, 7,866 and 33,359 ROIs that are uniquely detected by HATSEQ, MACS and FindPeaks respectively. (C) Distribution of the average read depth for all the detected regions using HATSEQ, MACS and FindPeaks.

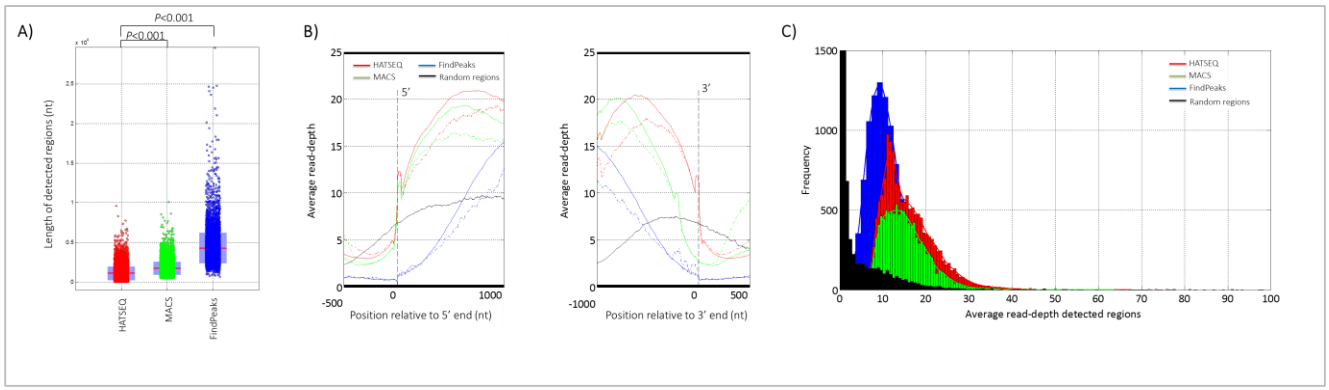


Figure 3. ROI statistics for the H3K4me experiment. Statistics for the detected ROIs by HATSEQ, MACS and FindPeaks (red, green and blue respectively) for the H3K4me experiment. (A) Boxplot illustrating the region length of the detected regions. (B) Average read depth across the ROI boundaries with respect to the 5' (left panel) and 3'end (right panel). The average read depths are calculated per nucleotide position after aligning the detected ROIs at their 5' and 3' ends, respectively. The solid line represents the alignment of the 14,616, 10,694 and 9,471 ROIs detected by HATSEQ, MACS and FindPeaks respectively. The dashed line represents the alignment of the 5,330, 1,305 and 601 ROIs that are uniquely detected by HATSEQ, MACS and FindPeaks respectively. (C) Distribution of the average read depth for all the detected regions using HATSEQ, MACS and FindPeaks.

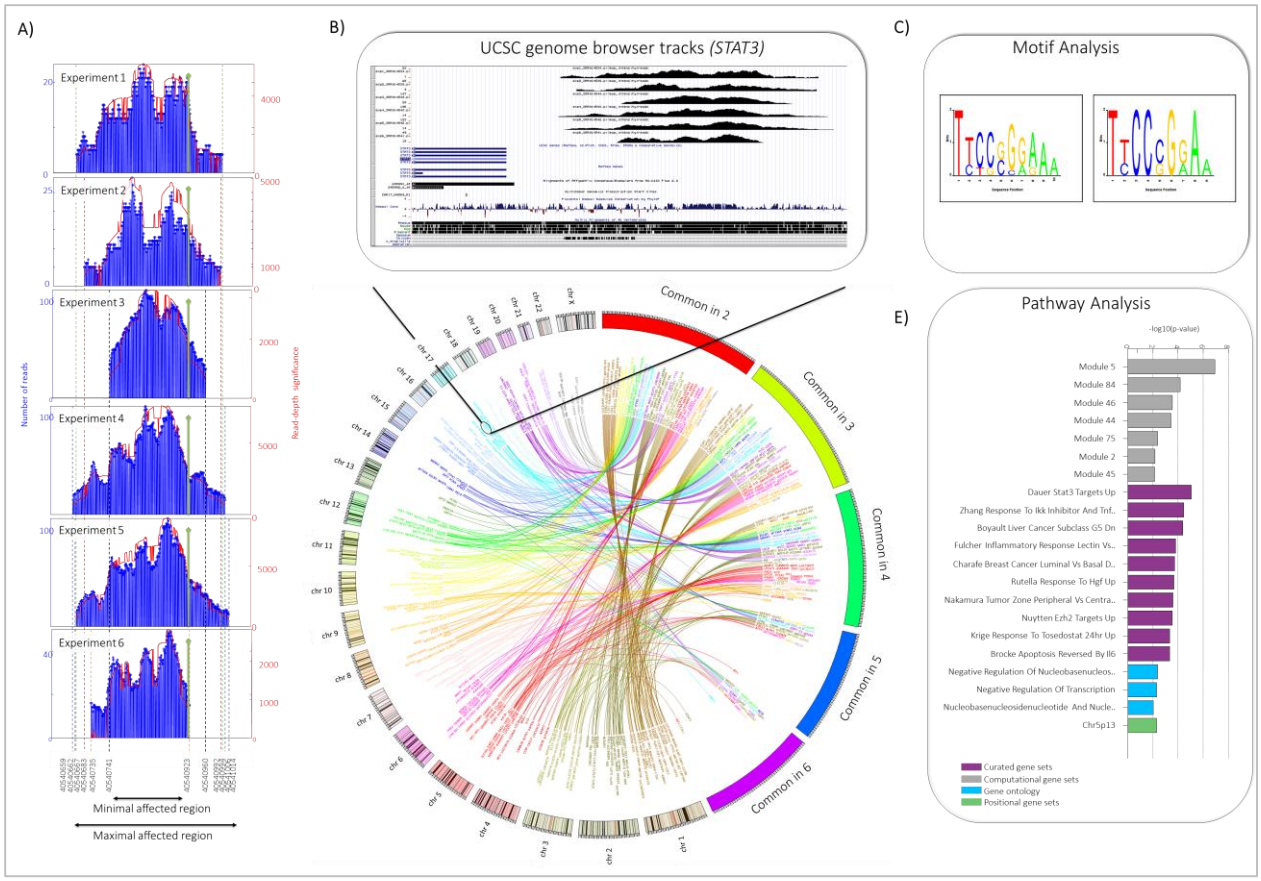


Figure 4. HATSEQ results for the STAT1 case study. The HATSEQ results of the STAT1 experiments using six interferon-γ (IFN-γ) stimulated human HeLa S3 cells compared to seven unstimulated human HeLa S3 cells. (A) Bar graph plot that illustrates a ROI that is detected in the promoter of *STAT3,* and seen across six experiments. The blue bars depicts the total number of reads per base pair position, indicated by the left y-axis. The red line illustrates the read depth significance score *Q(g)*, which reports how often reads were part of the statistically significant region, indicated by the right y-axis. The green bar illustrates the binding site of the expected STAT1 motif. (C) The top enriched motifs, among the 511 ROIs detected across two or more replicates. (D) Circos plot illustrating the genes, for which the closest detected ROI is detected among 2 or more experiments. A line connects selected genes, based on the chromosomal location with the number of experiments that a ROI is detected in. The colors indicate the chromosomal location of the genes. (E) Pathway analysis illustrates the enrichment for curated gene sets, computational gene sets, gene ontology and positional gene sets (with a maximum of ten gene sets in each category).

CHAPTER 4

Retroviral Integration Mutagenesis in Mice and Comparative Analysis in Human AML Identify Reduced *PTP4A3* Expression as a Prognostic Indicator

PLoS one

October 2011 | Volume 6 | Issue 10 | doi: 10.1371/journal.pone.0026537

Retroviral Integration Mutagenesis in Mice and Comparative Analysis in Human AML Identify Reduced *PTP4A3* Expression as a Prognostic Indicator

Reneé Beekman, Marijke Valkhof, Stefan J. Erkeland, Erdogan Taskesen, Veronika Rockova, Justine K. Peeters, Peter J. M. Valk, Bob Löwenberg and Ivo P. Touw

ABSTRACT

Acute myeloid leukemia (AML) results from multiple genetic and epigenetic aberrations, many of which remain unidentified. Frequent loss of large chromosomal regions marks haplo-insufficiency as one of the major mechanisms contributing to leukemogenesis. However, which haplo-insufficient genes (HIGs) are involved in leukemogenesis is largely unknown and powerful experimental strategies aimed at their identification are currently lacking. Here, we present a new approach to discover HIGs, using retroviral integration mutagenesis in mice in which methylated viral integration sites and neighbouring genes were identified. In total we mapped 6 genes which are flanked by methylated viral integration sites (mVIS). Three of these, i.e., Lrmp, Hcls1 and Prkrir, were upregulated and one, i.e., Ptp4a3, was downregulated in the affected tumor. Next, we investigated the role of PTP4A3 in human AML and we show that PTP4A3 expression is a negative prognostic indicator, independent of other prognostic parameters. In conclusion, our novel strategy has identified PTP4A3 to potentially have a role in AML, on one hand as a candidate HIG contributing to leukemogenesis in mice and on the other hand as a prognostic indicator in human AML.

INTRODUCTION

Acute myeloid leukemia (AML) is a complex disease driven by multiple cytogenetic abnormalities, such as inv(16), t(8;21), t(15;17), 3q abnormalities, deletions of (the q-arms) of chromosome 5 and 7 and by aberrant expression and/or mutations of genes e.g., *EVI1, FLT3, RAS, RUNX1, CKIT, WT1, CEBPA* and *NPM1*[119,120]. The frequent occurrence of chromosomal deletions suggests that haplo-insufficiencies contribute to the pathogenesis of AML. However, because deleted regions often harbor numerous genes, it remains difficult to pin point critical haplo-insufficient genes (HIGs) involved in the pathogenesis of AML. Gene expression profiling (GEP) focusing on downregulated genes could be informative, however differences in expression levels may relate to differentiation status of the AML blasts, rather than to mechanisms underlying leukemogenesis[121]. In addition, mapping of minimal affected regions in combination with GEP to identify HIGs often is cumbersome because these regions may still contain numerous genes and differences in their expression level may be subtle. Even in chromosomal regions frequently lost upon leukemic progression, e.g., the q-arm of chromosome 7, identification of critical HIGs remains difficult.

Retroviral insertion mutagenesis in mouse models has been used to discover novel genes involved in the development of different types of cancer[83,85,86]. Most of these genes have been classified as proto-oncogenes, owing to the fact that proviral integrations preferentially occur in 5′ promoter regions, supposedly leading to increased or sustained expression of flanking genes. Only a small minority of identified genes have been classified as tumor suppressor genes or HIGs, based on disruption of coding sequences by the proviral integration[122,123]. Gene therapy studies using murine leukemia virus (MLV)-based vectors have shown that epigenetic changes of long terminal repeats (LTRs) of integrated proviruses often result in silencing of therapeutic genes[124,125], and that preventing methylation of the CpG islands within LTRs overcomes this problem[126]. Based on these observations, we hypothesized that methylation of viral sequences not only results in silencing of retroviral genes themselves but may also affect host genes located proximal to proviral integrations. Methylated LTRs located in proximity of promoter regions may thus identify genes that are deregulated leading to haplo-insufficiency.

To discover potential HIGs relevant for human AML, we used murine leukemia samples induced by Graffi 1.4 Murine Leukemia Virus (Gr1.4 MLV), classified as mixed lineage or myeloid leukemias by immunophenotyping[83,127]. By methylation specific PCR (MSP) and methylated DNA immunoprecipitation (MeDIP)[128] we observed an extensive variation in the level of DNA methylated proviral integrations in these tumors. We designed a strategy to map methylated proviral integrations by combining MeDIP, inverse PCR (iPCR) and promoter array hybridization. We identified 6 genes to be flanked by methylated viral integration sites (mVIS), of which *Lrmp*, *Hcls1* and *Prkrir* were transcriptionally upregulated and *Ptp4a3* was transcriptionally downregulated. Further studies in human AML samples revealed a negative prognostic value of *PTP4A3* expression levels, independent of other prognostic indicators. In conclusion, by mapping DNA methylated viral integration sites in murine leukemias induced by retroviral integration mutagenesis followed by comparative analysis in human AML, we identified *PTP4A3* not only as a candidate HIG contributing to leukemogenesis in mice but also as an independent prognostic indicator in human AML.

RESULTS

Viral integrations sites of the Graffi1.4 MuLV are subject to DNA-methylation

In this study murine leukemia samples induced by Gr1.4 MLV were analysed[83]. First, a methylation specific PCR (MSP) was performed to determine the level of DNA-methylation of the Gr1.4 MLV LTRs. To this end, amplification products from methylated LTRs were quantified with quantitative PCR (qPCR) and corrected for total LTRs in these samples (Figure 1A). A considerable variation in LTR methylation was seen between different tumors (data not shown). Based on these methylation levels, leukemia samples were divided into 4 methylation categories of equal sample size (1 = highest LTR methylation level, 4 = lowest LTR methylation level).

Subsequently, MeDIP was used on a subset of samples to enrich for methylated LTRs and flanking genomic regions. As a control, genomic DNA of normal bone marrow, spleen and liver was used. MeDIP enrichment relative to input

levels was determined for the LTR, the non-methylated actin B locus (*ActB*) and the hemi-methylated imprinting control region 1 (ICR1) of *H19*. As expected, *H19* enrichment scores were high and *ActB* enrichment scores were low in all categories (Figure 1B). Additionally, samples in the highest methylation category showed a significantly higher LTR enrichment after MeDIP compared to the samples in other categories (P-value <0.001), confirming the specificity of the MSP (Figure 1B).

Ptp4a3 is flanked by a methylated viral integration site and is transcriptionally downregulated

Genes located near methylated viral integration sites (mVIS) may be downregulated due to the proximity of a methylated regulatory sequence, and, their transcriptional downregulation may contribute to murine leukemogenesis. Therefore, after showing that a proportion of viral integration sites are subject to DNA-methylation, we set out to identify genes flanking these viral integration sites. To this end, iPCR, to amplify regions flanking viral integration sites, and MeDIP, to enrich for DNA methylated fragments, were combined to amplify regions flanking mVIS (Figure 2). Amplified fragments of 6 tumor samples were hybridized to Murine 1.0 R promoter arrays and, using Hypergeometric analysis of tiling-arrays (HAT)[89], 15 amplified regions were mapped in these tumors (Table S1). Eight of these integrations were validated by directed PCR followed by Sanger sequencing (Figure 3, Table S1). Because MLVs tend to integrate within 10 kb around the transcriptional-start-site[129], the nearest genes within 10 kb downstream of these 8 mVIS were determined (Figure 3, Table S1).

To support that regions identified in this way were indeed flanked by methylated LTRs, we performed a methylation sensitive digestion followed by directed PCR. Using this approach, only viral integration sites flanked by methylated LTRs could be amplified (Figure 4A), as was the case for 6 out of 8 identified integrations (Figure 4B, Table S1). Subsequently, expression levels of genes flanking these mVIS were quantified by qPCR and compared to normal bone marrow expression levels. Unfortunately, RNA of tumor 1 was lacking, therefore this analysis could not be performed for *Taf12* and *Ranbp3.* Of the other 4 genes, *Ptp4a3* expression was 2–3 fold reduced in the respective tumor (Figure 4C, Table S1).

Ptp4a3 is an independent prognostic factor in human AML

The human orthologue of murine *Ptp4a3*, i.e., *PTP4A3*, was further studied in human AML. Transcript levels of *PTP4A3* were assessed in 454 AML samples, diagnosed under the age of 60, profiled using the HGU133 2.0 plus gene expression arrays[62]. *PTP4A3* expression values are represented by 2 probesets with a high correlation (Pearson correlation coefficient = 0.90). Survival analysis with these probesets gave similar results; all results shown are based on expression levels of probeset 206574_s_at. *PTP4A3* expression levels were negatively correlated with prognostic outcome both for overall survival (OS, P-value <0.0001, hazard ratio = 1.269) and event-free survival (EFS, P-value <0.0001, hazard ratio = 1.261). Kaplan-Meier curves are shown in Figure 5. A permutation test predicted a probability of 0.0036 for a random gene locus to be a significant prognostic indicator with a P-value <0.0001 for both OS and EFS. Multivariate analysis showed that the negative correlation of *PTP4A3* expression with event-free survival was

independent of other prognostic parameters, i.e., age, white blood cell count, cytogenetic risk, *CEBPA* mutation status and *NPM1*$^{+}$*FLT3ITD*$^{-}$ status (Table 1).

DISCUSSION

We designed a strategy to identify candidate HIGs in AML using retroviral integration mutagenesis, by mapping DNA methylated proviral integrations. By using HAT[89], we deliberately aimed at detecting integrations present in the majority of the leukemic cells, which are most likely involved in the early phase of leukemogenesis. At the same time, integrations present in subclones that contribute to later stages of leukemic progression will be missed using this approach. We identified 6 genes that are flanked by methylated viral integrations. Expression analysis showed that *Lrmp* (lymphoid-restricted membrane protein), *Hcls1* (hematopoietic cell specific Lyn substrate 1) and *Prkrir* (protein-kinase, interferon-inducible double stranded RNA dependent inhibitor, repressor of (P58 repressor)) were upregulated and *Ptp4a3* (protein tyrosine phosphatase type IVA), a phosphatase also known as *Prl3* (phosphatase of regenerating liver 3) was downregulated in the respective murine tumor. These results indicate that a flanking methylated viral integration site does not necessarily lead to transcriptional repression. As 1 out of 4 genes flanked by a mVIS was transcriptionally downregulated and expression of the 2 other genes could not be investigated, the efficiency to detect potential HIGs by identifying mVIS would approximately be 17–25%. However, the number of analysed tumors is too small to allow an accurate estimation of the efficiency.

Ptp4a3 expression is controlled by p53 induced after DNA damage in mouse embryonic fibroblasts (MEFs) and its activity is involved in inducing a G1 cell cycle arrest in these cells[130]. Surprisingly however, the same study also demonstrated a cell cycle arrest upon reduction of *PTP4A3* expression[130]. Apparently, depending on expression level dosage, *PTP4A3* may have both positive and negative effects on cell cycle regulation. Hence, *PTP4A3* haplo-insufficiency, but not its complete loss, may lead to an impairment of cell cycle arrest after DNA damage. Dosage effects of *PTP4A3* expression in relation to cellular responses may be more complex, particularly in cancer cells. For example, in carcinoma cell lines *PTP4A3* expression may lead to downregulation of p53[131] and it is variably induced by γ-irradiation[132]. Finally, high *PTP4A3* expression has been linked to increased tumor aggressiveness in different types of solid tumors, e.g., melanoma, gastric cancer, colon cancer, hepatocellular carcinoma and breast cancer[133-138], possibly because high *PTP4A3* expression leads to increased epithelial-mesenchymal transition[138].

The role of *PTP4A3* in hematopoietic malignancies has not been studied as extensively as in carcinoma. Only a few studies report differences in expression levels of *PTP4A3* in ALL and myeloma subgroups, based on gene expression profiling[139-141]. Interestingly however, in a recent study, *PTP4A3* has been proposed to have a role in drug-resistance in AMLs with internal tandem duplication of *FLT3* (*FLT3ITD*)[142]. This finding, together with the observation that high *PTP4A3* expression negatively correlates with prognostic outcome, indicates that PTP4A3 might be a potential therapeutic target in AML.

In conclusion, using a retroviral mutagenesis screen in which we enriched for DNA methylated viral integration sites we identified *PTP4A3* as a potential haplo-insufficient gene with an independent prognostic value in human de novo AML. Challenges for the future are to determine the dose-effect of *PTP4A3* expression in myeloid development and to extend the screens to additional myeloid neoplasms, e.g., myelodysplasia, therapy-related AML, AML secondary to bone marrow failure and myeloproliferative disorders.

MATERIALS AND METHODS

Ethics statement

For this study no novel murine leukemias were generated, all experiments described were performed on material generated in a previous study[83]. All animal procedures for the use of control bone marrow fractions were approved by the animal care and use committee of the Erasmus MC (approval # 119-10-05).

All human cell samples were obtained after written informed consent and stored anonymously in a biobank. The study was performed under the permission of the Institutional Review Board of the Erasmus MC, registration number MEC-2008-387.

Mouse leukemia and normal cell samples

DNA and RNA samples from a previously generated panel of Gr1.4-induced leukemia's[83], and control samples (bone marrow, spleen, liver) from normal FVB/N mice were used.

Methylation specific PCR

Primer and probe sequences are shown in Table S2. Two µg of genomic DNA was treated with bisulphite using the EZ DNA-methylation kit according to the manufacturer's protocol (Zymo research, Orange, CA, USA). LTRs were amplified with bsLTRfw and bsLTRrv using 1 µL out of 10 µL of bisulphite-treated DNA. Cycling conditions were 30″ at 94°C, 30″ at 50°C and 1′ at 72°C for 10 cycles in a total volume of 50 µL. Two µL was used in a nested qPCR (Figure 1A) using MN-LTR-fw×MS-LTR-rv/MN-LTR-rv (MN = methylation neutral, MS = methylation specific). Cycling conditions were 15″ at 94°C, 30″ at 57°C and 30″ at 60°C for 45 cycles. Amplified LTRs, methylated and unmethylated, were quantified using a methylation neutral probe (probe-MN, Sigma-Aldrich, Zwijndrecht, The Netherlands). Delta cycle threshold values (dCt), representing the number of methylated LTRs as a fraction of total LTRs, were calculated as follows: dCt = Ct(Methylated LTRs)-Ct(All LTRs) = Ct(MN-LTR-fw×MS-LTR-rv) – Ct(MN-LTR-fw×MN-LTR-rv). PCRs were performed in duplicate and mean dCt values were calculated.

MeDIP

Ten µg genomic DNA was digested overnight with 100 U of DpnII (New England Biolabs, Ipswich, MA, USA). Four µg digested DNA was denatured for 10′ at 95°C and incubated with either 2.5 µg anti-5-methylcytidine (BI-MECY-1000,

Eurogentec, Liège, Belgium) or mouse pre-immune IgG (Sigma-Aldrich, Zwijndrecht, The Netherlands) in 500 µL IP-buffer (PBS with 0.05% Triton X-100) for 2 hrs at 4°C, followed by incubation with 30 µL of washed beads (M-280 sheep-anti-mouse IgG, Invitrogen, San Diego, CA, USA) for 2 hrs at 4°C. Beads were washed 3 times with 700 µL IP-buffer. As a 10% input reference, 400 ng digested DNA not subjected to MeDIP was used. Beads and the 10% input reference DNA were resuspended in 100 µL IP-buffer and incubated for 3 hrs at 50°C after adding 20 µg proteinase K (Roche, Basel, Switzerland). Supernatants, containing immunoprecipitated DNA, and the input DNA were purified using the MinElute Reaction Cleanup Kit (Qiagen, Hilden, Germany) and were eluted in 40 µL elution buffer. Two µL immunoprecipitated DNA was used to amplify the imprinting control region 1 (ICR1) of *H19* with *H19*ICR1fw × *H19*ICR1rv, *ActB* with *ActB*fw × *ActB*rv and the LTR with LTRfw × LTRrv using (q)PCR. Primer sequences are shown in Table S2. Cycling conditions were 30″ at 95°C, 30″ at 58°C and 45″ at 72°C for 30 cycles (PCR) or 15″ at 94°C, 30″ at 59°C and 30″ at 60°C for 45 cycles (qPCR). Amplification products were analysed using gel electrophoresis (PCR) or quantified (qPCR) using SYBRgreen Master mix (Applied Biosystems, Foster City, CA, USA).

Inverse PCR

Primer sequences are shown in Table S2. Six murine leukemias with high LTR enrichment (more than 10% of input) and low *ActB* enrichment (less than 10% of input) were selected for inverse PCR. Eight µL MeDIP-DNA was denatured for 3′ at 95°C, renatured by a temperature decrease of 0.1°C/sec to 20°C, and ligated for 45′ at room temperature using a rapid DNA ligation kit (Roche, Basel, Schwitzerland). Two µL out of 20 µl ligated product was amplified with primers mL1 and mL2, followed by a nested PCR with primers mL1N and mL2N using 2 µL of the first PCR product. Cycling conditions were 30″ at 95°C, 30″ at 60°C (first PCR) or 56°C (nested PCR) and 3′ at 72°C for 30 cycles. In the nested PCR 10 mM dCTP, dATP, dGTP, 8 mM dTTP and 2 mM dUTPs were used.

Promoter array hybridization

PCR products of 10 nested PCR reactions were purified with a PCR purification kit (Qiagen, Hilden, Germany) and pooled. A total of 7.5 µg of these amplified fragments was fragmented and labeled using the GeneChip WT Double stranded DNA terminal labeling kit (Affymetrix, Santa Clara, CA, USA). Fragmentation to 66 bp was checked on a Bioanalyser (Agilent, Santa Clara, CA). Labeled DNA was hybridized to mouse promoter 1.0R arrays (Affymetrix, Santa Clara, CA, USA) for 16 hrs at 45°C. Arrays were washed with the FS_450_0001 protocol using the Fluidics Station 450 (Affymetrix, Santa Clara, CA, USA), followed by scanning. Probe values were normalized with model-based analysis of tiling-arrays (MAT)[74] and mVIS were determined using hypergeometric analysis of tiling-arrays (HAT)[89], both for HAT and MAT default settings were used. Genes located nearby amplified regions were identified using UCSC (assembly mm8, Feb. 2006).

Directed PCR and Sanger sequencing

Primers are shown in Table S2; amplification of the integration site was performed with VIS(*corresponding gene*) × LTRfw2, for *Lrmp* a nested PCR was performed with VIS(*Lrmp_nested*) × LTRfw. As input, 200 ng of the corresponding tumor DNA was used; cycling conditions were 30″ at 95°C, 30″ at 58°C and 45″ at 72°C for 30 cycles. Products were purified using the Multiscreen HTS 66-well filtration system (Millipore, Billerica, MA, USA). Sanger sequencing was performed with primer LTRfw according to the manufacturer's protocol (Applied Biosystems, Foster City, CA, USA).

Methylation sensitive restriction analysis

Primers are shown in Table S2. Two and a half µg of tumor DNA was digested with 25 U of BstU1 (New England Biolabs, Ipswich, MA, USA) o/n at 60°C, purified using the Multiscreen HTS 66-well filtration system (Millipore, Billerica, MA, USA), eluted in 30 µl and diluted to 50 ng/µl. Amplification of the integration site was performed as described under directed PCR and Sanger sequencing, with 100 instead of 200 ng input of DNA. As controls *H19* ICR1 (*H19*ICR1fw × *H19*ICR1rv) and *ActB* (*ActB*fw ×*ActB*rv) were amplified. Cycling conditions were 30″ at 95°C, 30″ at 58°C and 45″ at 72°C for 30 cycles. Amplification products were analysed using gel electrophoresis.

RNA isolation, cDNA preparation and qPCR

RNA of murine samples was isolated using Trizol (Invitrogen, San Diego, CA) according to the manufacturer's protocol. One µg of RNA was used for cDNA preparation, using SuperScript II Reverse Transcriptase (Invitrogen, San Diego, CA) according to the manufacturer's protocol. One µl cDNA was used as input for the qPCR. Genes of interest were amplified with their respective forward and reverse primers (Table S2), as an input control, TATA box binding protein (*Tbp*) was analysed. Cycling conditions were 3″ at 95°C and 30″ at 60°C for 45 cycles. Amplification products were quantified using Fast SYBRgreen Master mix (Applied Biosystems, Foster City, CA, USA). Expression levels relative to *Tbp* were calculated.

Survival analysis human AML samples

Purified AML blasts were obtained following informed consent as described[18]. Gene expression profiles of 454 de novo AML patients under the age of 60 were used for this analysis[62]. Expression levels were MAS5 normalised (Scaling factor 100), values <30 were set at 30, followed by log2 transformation.

For *Ptp4a3*, univariate and multivariate survival analyses were performed using expression levels of probesets 206574_s_at or 209695_at in a Cox regression model. In the multivariate analysis age, white blood cell count, cytogenetic risk group, *NPM1*$^{+}$*FLT3ITD*$^{-}$ status and *CEBPA* mutation status were used as additional prognostic parameters. We recognised the following cytogenetic risk groups: favorable = t(15;17), inv(16) and t(8;21), unfavorable = t(3;3), inv(3), −7/7q-, −5/5q-, complex karyotype, t(11q23) except t(9;11), t(9;22) and t(6;9), intermediate = all other cases with known cytogenetics. Kaplan-meier graphs were generated by dividing the AML cohort in 2 groups of equal sample size based on *PTP4A*3 expression of probe 206574_s_at. Analyses were performed in SPSS (version 17, SPPS Inc, Chicago, IL).

For the permutation test, all probesets with an annotated gene symbol (based on HG-U133_Plus_2.na32.annot.csv, Affymetrix, Santa Clara, CA, USA) were selected. Next probesets with expression levels <30 in all 454 patients were discarded, leaving a total of 40720 probesets. The permutation test was performed by randomly selecting 6 probesets (representing 6 mVIS), followed by randomly selecting 1 out of these 6 probesets (representing 1 downregulated gene). For this probeset a univariate Cox regression analysis was performed for overall survival (OS) and event-free survival (EFS). A P-value of <0.0001 (as observed for *PTP4A3*) was considered significant. This analysis was repeated 100.000 times, followed by calculating the frequency, i.e., probability, of observing a significant P-value for both OS and EFS. Analyses were performed in Matlab (version 2008b, Mathworks, Natick, MA).

ACKNOWLEDGMENTS

The authors thank M. Sanders for advice concerning the permutation test.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: RB MV SJE IPT. Performed the experiments: RB MV. Analyzed the data: RB ET VR IPT. Contributed reagents/materials/analysis tools: ET PJMV JKP BL IPT. Wrote the paper: RB IPT.

FIGURE LEGENDS

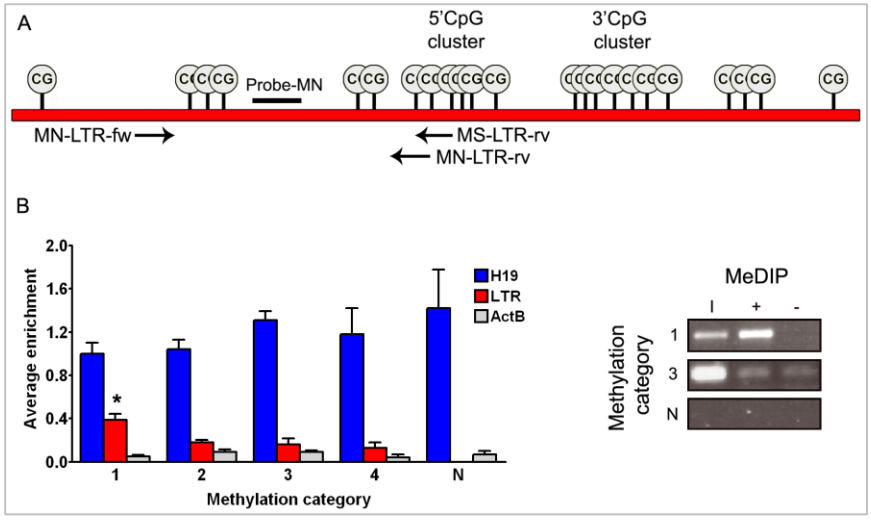


Figure 1. LTR methylation analysis. (A) Overview of the methylation specific PCR (MSP) approach. Depicted is a schematic representation of the Gr1.4 MuLV LTR, containing 23 CpGs. The MSP was performed, after bisulphite treatment, with a methylation neutral forward primer (MN-LTR-fw), and a methylation specific (MS-LTR-rv) or neutral (MN-LTR-rv) reverse primer. Amplification products were quantified using methylation neutral probe-MN. Tumor samples were divided into 4 equal groups based on the methylation status of their LTRs (1 = highest LTR methylation

level, 4 = lowest LTR methylation level). (B) *Left panel.* For *H19*, the LTR and *ActB*, average enrichment after MeDIP compared to input levels were calculated for each MSP-defined methylation category as well as for normal bone marrow, spleen and liver (N). In category 1 to 4 respectively 18, 15, 4 and 3 samples were analysed; error bars indicate standard deviations. P-values were calculated using a Wilcoxon test; *significantly higher than other categories, P-value <0.001. *Right panel*. Example of LTR enrichment after MeDIP (I = input, + = IP with anti-5-methylcytidine, – = IP with pre-immune serum IgG, 1 and 3 = methylation categories, N = normal spleen).

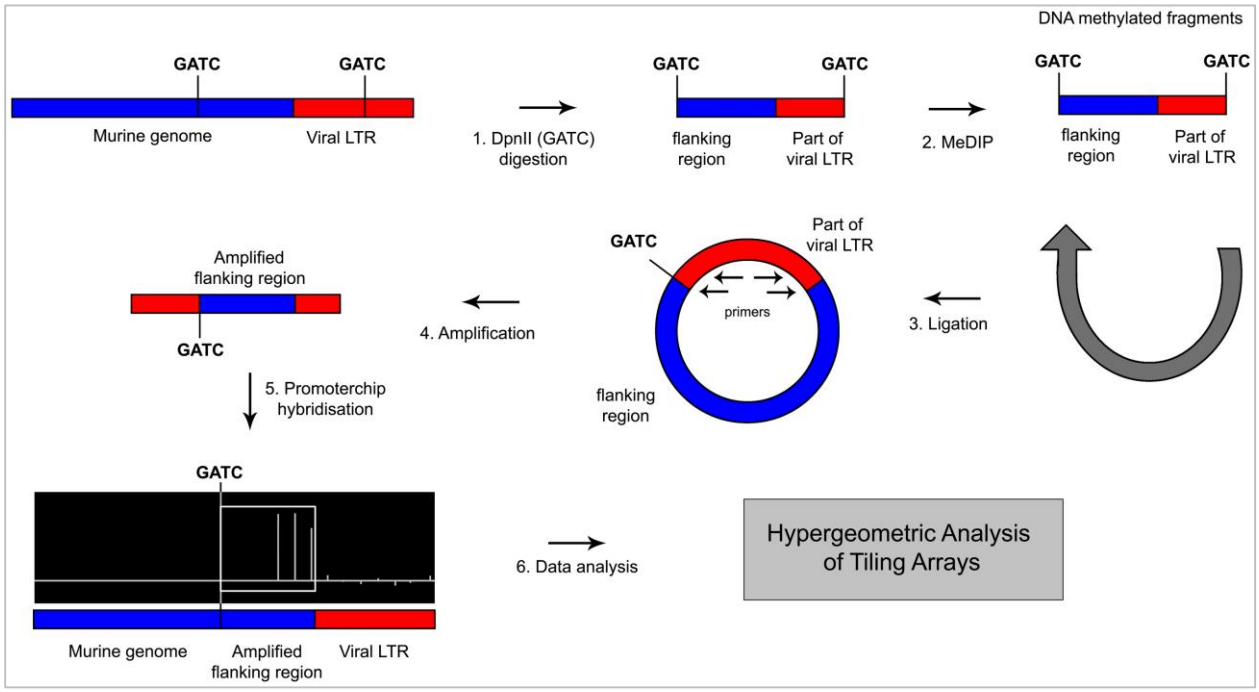


Figure 2. Identification of mVIS. Strategy outline for identification of regions flanking DNA methylated viral integration sites (mVIS) within murine leukemias. Genomic DNA was digested with DpnII (step 1), followed by methylated DNA immunoprecipitation (MeDIP, step 2). MeDIP enriched fragments were ligated (step 3) and amplified using primers within the LTR (step 4). These fragments were hybridized on a DNA promoter array (step 5). Hypergeometric Analysis of Tiling-arrays (HAT) was used to identify regions flanking mVIS (step 6).

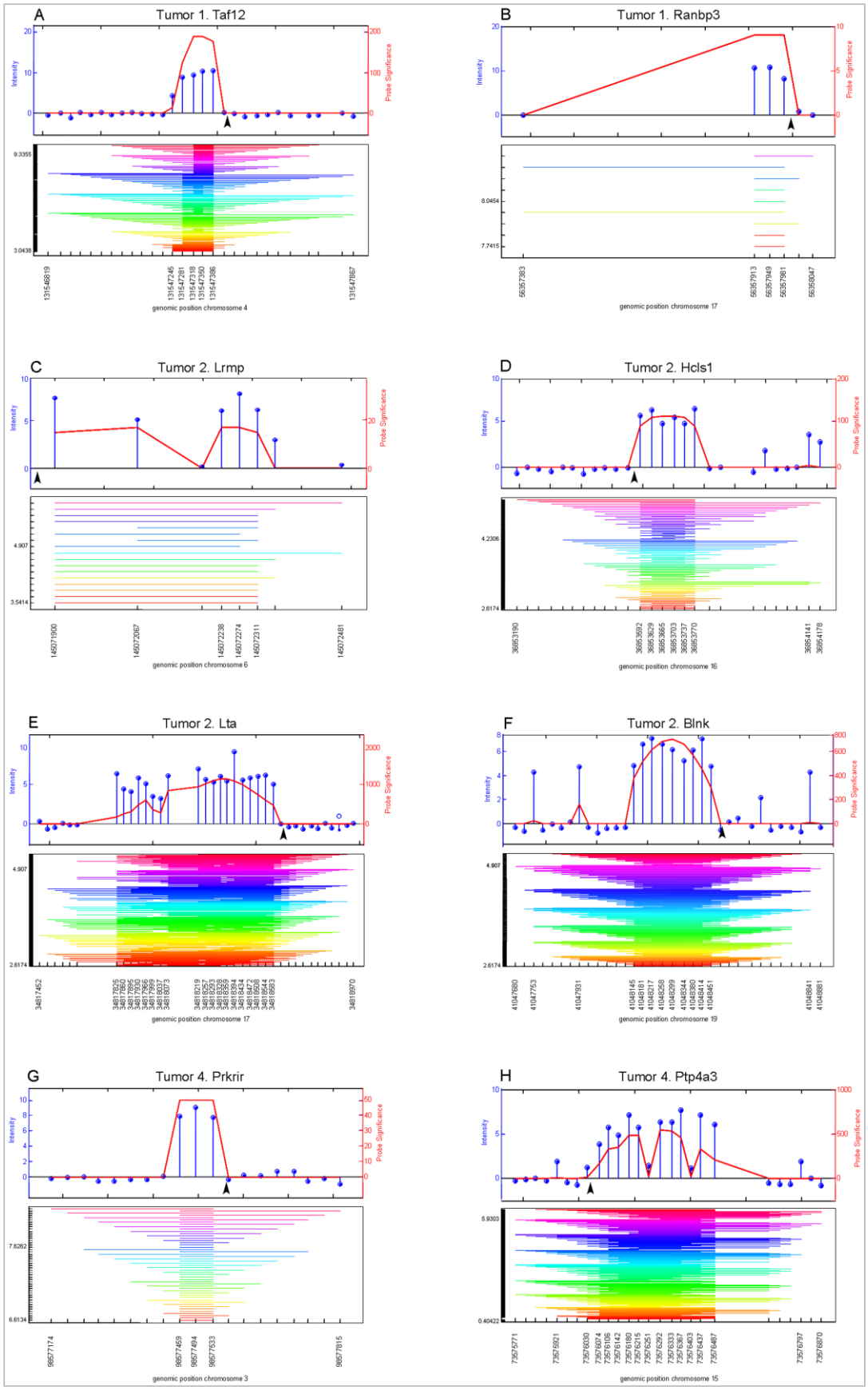

A
Tumor 1. Taf12
Intensity
Probe Significance
genomic position chromosome 4
B
Tumor 1. Ranbp3
genomic position chromosome 17
C
Tumor 2. Lrmp
genomic position chromosome 6
D
Tumor 2. Hcls1
genomic position chromosome 16
E
Tumor 2. Lta
genomic position chromosome 17
F
Tumor 2. Blnk
genomic position chromosome 19
G
Tumor 4. Prkrir
genomic position chromosome 3
H
Tumor 4. Ptp4a3
genomic position chromosome 15

Figure 3. Identified viral integration sites. Eight viral integrations identified with HAT could be confirmed with directed PCR and Sanger sequencing (see Table S1 for further details). The graphical output of HAT is represented in graph A–H. Above each graph, the tumor in which the integration was identified as well as the nearby located gene are indicated. The upper panel of each graph shows normalized intensities of the different probes (blue lollipops) on the mouse promoter 1.0R arrays and their significance (in red) as calculated with HAT. The black arrowhead indicates the exact position of the proviral integration, as determined by directed PCR followed by Sanger sequencing. In the lower panel the lowest and highest probe intensity threshold with a significant outcome are given on the left. The stripes indicate significantly enriched regions at different probe intensity thresholds, calculated with HAT, which are merged into the final viral integration site. Below each graph, the genomic position is indicated (assembly mm8, February 2006).

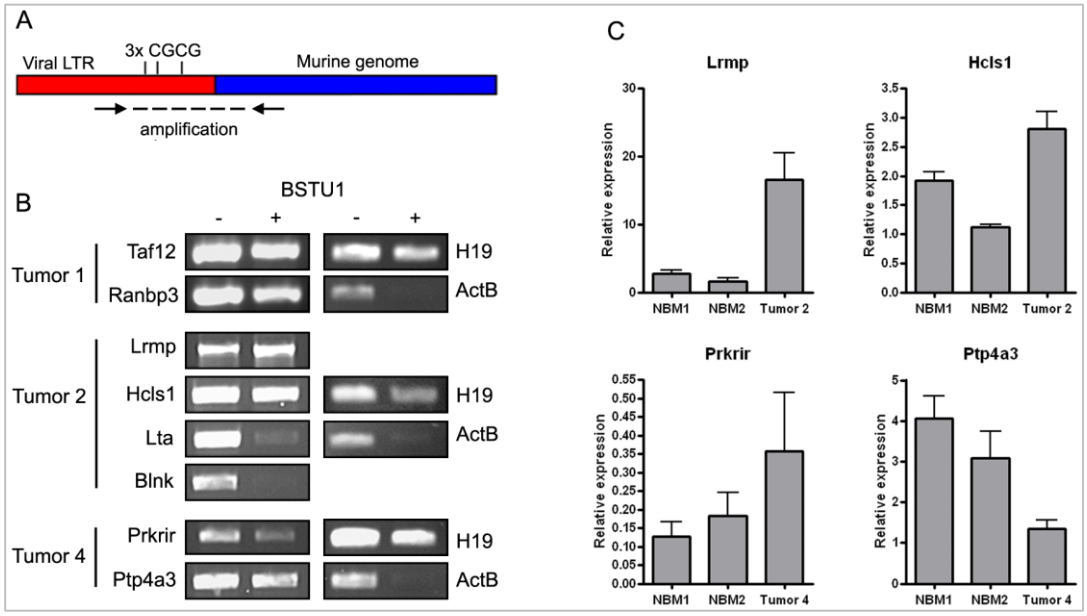


Figure 4. Methylation sensitive restriction analysis of viral integration sites and expression of nearby located genes. (A) Schematic overview of the methylation specific restriction approach. Genomic DNA was digested with BstU1 (CGCG, blocked by DNA-methylation), followed by mVIS amplification with primers as indicated by arrows. If the flanking LTR is methylated, mVIS amplification is unaffected upon BstU1 digestion. (B) All 8 identified viral integration sites, identified in tumor 1, 2 and 4, were amplified before (–) and after (+) BstU1 digestion. As controls, *H19* (hemi-methylated) and *ActB* (unmethylated), both containing 2 BstU1 digestion sites, were analysed in each tumor. (C) Expression levels of 4 genes flanked by methylated viral integration sites were determined by qPCR in the respective tumors. Expression levels relative to housekeeping gene *Tbp* are shown; error bars indicate standard deviations. NBM=normal bone marrow.

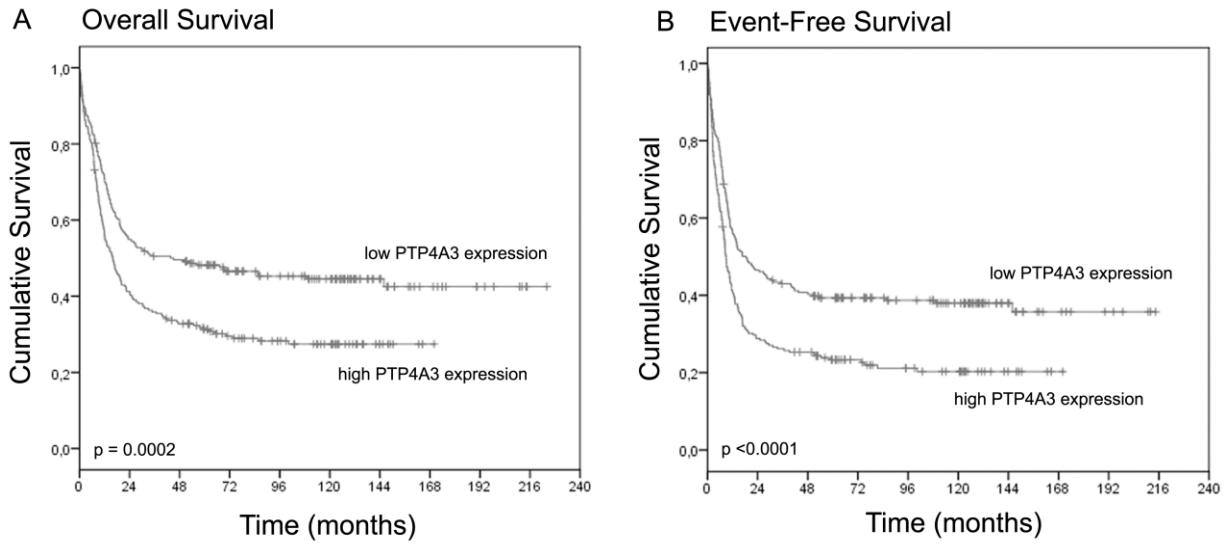


Figure 5. Survival analysis. A cohort of 454 de novo AML cases diagnosed under the age of 60 was divided into 2 groups of equal size based on MAS5 normalised expression of *PTP4A3* (probe 206574_s_at). Overall survival (A) and event-free survival (B) were analysed. P-values were calculated with a log rank test.

Risk factor	Overall Survival HR (95% CI)	*P*-value	Event-Free HR (95% CI)	*P*-value
PTP4A3 expression	1.112 (0.995–1.243)	0.061	1.131 (1.019–1.255)	0.021*
Age (decades)	1.134 (1.024–1.256)	0.016*	1.068 (0.969–1.177)	0.186
WBC$^{\infty}$	1.373 (1.063–1.773)	0.015*	1.296 (1.020–1.648)	0.034*
Favorable cytogenetic risk†	0.376 (0.257–0.548)	<0.0001*	0.469 (0.335–0.658)	<0.0001*
Unfavorable cytogenetic risk†	1.432 (1.059–1.935)	0.020*	1.507 (1.124–2.020)	0.006*
NPM1^{+}*FLT3*$^{ITD-}$ ‡	0.473 (0.317–0.705)	0.0002*	0.578 (0.398–0.839)	0.004*
CEBPA double mutant$^{\$}$	0.591 (0.418–0.836)	0.003*	0.560 (0.384–0.815)	0.002*

Table 1. Multivariate survival analysis.

Multivariate analysis in 454 de novo AML patients under the age of 60.
$^{\infty}$WBC higher than 20×10^9/L versus lower than 20×10^9/L,
†compared to intermediate cytogenetic risk,
‡compared to no NPM1^{+}FLT3^{ITD-},
$^{\$}$compared to no CEBPA double mutation.
*Statistically significant. HR: hazard ratio, CI: confidence interval, WBC: white blood cell count, FLT3^{ITD}: internal tandem duplication of FLT3.

SUPPORTING MATERIAL

Tumor name	Chromosome	Region start (mm8)	Region stop (mm8)	Region Size	Confirmed by directed PCR and Sanger sequencing	Exact position integration determined with Sanger sequencing (mm8)	Nearest gene	Distance from gene	Flanked by methylated LTR as confirmed by methylated restriction analysis	Down regulated in tumor
Tumor1	chr4	131546819	131547867	1048	Yes	131547433	*Taf12*	1st intron	Yes	Not determined due to lack of material
Tumor1	chr8	75652987	75653413	426	No	-	-	-	-	-
Tumor1	chr17	56357383	56358047	664	Yes	56357998	*Ranbp3*	379 bp upstream	Yes	Not determined due to lack of material
Tumor2	chr1	133792492	133792605	113	No	-	-	1st intron	-	-
Tumor2	chr2	164274288	164274798	510	No	-	-	-	-	-
Tumor2	chr6	145071900	145072481	581	Yes	145071861	*Lrmp*	7037 bp upstream	Yes	No
Tumor2	chr16	36853190	36854178	988	Yes	36853575	*Hcls1*	647 bp upstream	Yes	No
Tumor2	chr17	34817452	34818970	1518	Yes	34818633	*Lta*	5230 bp upstream	No	-
Tumor2	chr19	41047680	41048881	1201	Yes	41048491	*Blnk*	645 bp upstream	No	-
Tumor4	chr2	85563399	85563837	438	No	-	-	-	-	-
Tumor4	chr7	98577174	98577815	641	Yes	98577566	*Prkrir*	989 bp upstream	Yes	No
Tumor4	chr15	73575771	73576870	1099	Yes	73576038	*Ptp4a3*	1st intron	Yes	Yes
Tumor5	chr2	164334437	164334992	555	No	-	-	-	-	-
Tumor5	chr3	20436615	20436752	137	No	-	-	-	-	-
Tumor6	chr15	73575994	73576487	493	No	-	-	-	-	-

Table S1. Retroviral integrations. Retroviral integrations identified with HAT are listed. For each integration the murine tumor and the genomic position are indicated as well as whether the integration could be confirmed with directed PCR and Sanger sequencing. For all integrations that could be confirmed, nearby located genes are given, their distance to the retroviral integration and whether the flanking viral integration was DNA methylated as analysed by methylation sensitive restriction analysis. Finally, for the 6 genes with a flanking DNA methylated viral integration site is indicated if they were downregulated in the respective tumor.

Name	Sequence
bsLTRfw	GAGAAATAGGGAAGTTTAGATTAA
bsLTRrv	CCCAAAATAAACAATCAATCAATC
MN-LTR-fw	GGTTAAATAGGATATTTGTGGTGAGTAG
MN-LTR-rv	AACGAACTAATTAATTAATTCAAATAAAAC
MS-LTR-rv	CGAACAAAAACGAAAAACGAA
Probe-MN	FAM-AAACCATATCTAAAAACCATCTATTCTTACCCCC-TAMRA
H19 ICR1fw	ACATTCACACGAGCATCCAGG
H19 ICR1rv	GCTCTTTAGGTTTGGCGCAAT
ActB fw	AGCCAACTTTACGCCTAGCGT
ActB rv	TCTCAAGATGGACCTAATACG
LTRfw	AAAGACCTGAAACGACCTTGC
LTRrv	AAGGACCAGCGAGACCACG
mL1	CAACCTGGAAACATCTGATGG
mL2	CCCAAGAACCCTTACTCGGC
mL1N	CTTGAAACTGCTGAGGGTTA
mL2N	AGTCCTCCGATAGACTGTGTC
LTRfw2	CCAGGTTGCCCCAAAGACCTG
VIS(*Taf12*)	CAAGATCCGGGCTTTCAGAC
VIS(*Ranbp3*)	GACCAGGCTGCTCTCAAACG
VIS(*Lrmp*)	GGACACTACACTCATATTTG
VIS(*Lrmp_nested*)	GTGTGCTATGGGTAATTCAG
VIS(*Hcls1*)	TTCTCCTCCTTGCTTTCTGC
VIS(*Lta*)	CTAGGAGTCTTGTGCATCGTC
VIS(*Blnk*)	GAGGACAAGCCTAGTGATTTC
VIS(*Prkrir*)	CTGCTTGTTCACACAAAGTC
VIS(*Ptp4a3*)	CAGCCTCCTCTAGCAGTATC
Tbp fw	GCTGACCCACCAGCAGTTCAGTA
Tbp rv	AAGGAGAACAATTCTGGGTTTGA
Lrmp fw	CACAAGGCGAAGAGGCAGTG
Lrmp rv	GTGCTCTGTTGGCTCTTCTG
Hcls1 fw	CCCTTCTCTGTCCTACCAAG
Hcls1 rv	CCTTCATCCACCATCTCAAT
Prkrir fw	CTTACCAGTCATTTGAACAAC
Prkrir rv	CTTCAAGGGTTAAAGGCAGC
Ptp4a3 fw	CCATCCAGTTCATCCGACAG
Ptp4a3 rv	GACACAGATGTAATGAGGTAC

Table S2. Primers and probes.

CHAPTER 5

HAT: A Novel Statistical Approach to Discover Functional Regions in the Genome

Springer Series

May 2013 | Volume 1067 | Issue 3 | doi: 10.1007/978-1-62703-607-8

HAT: A Novel Statistical Approach to Discover Functional Regions in the Genome

Erdogan Taskesen, Bas J. Wouters and Ruud Delwel

ABSTRACT

Tiling-arrays are useful for exploring local functions of regions of the genome in an unbiased fashion. The exact determination of those genomic regions based on tiling-array data, e.g., generated by means of hybridization with immunoprecipitated DNA-fragments to the arrays is a challenge. Many different statistical methodologies have been developed to find biological relevant regions-of-interest (ROI) by using the quantitative signal intensity of each probe. We previously developed a method called Hypergeometric Analysis of Tiling-arrays (HAT) for the analysis of tiling-array data, but it is developed such that it can also be used to study data derived by genome wide deep sequencing approaches. Here we applied HAT to analyze two publicly available Tiling-array data sets. After the detection of statistically significant ROI, these are often used in additional analysis for hypothesis testing. We therefore discuss, by using the results of the tiling-array experiment, pathway and motif analyses.

INTRODUCTION

Tiling-arrays are a subtype of microarrays which are designed with probes that cover contiguous regions of a genome. The locations of probes do not necessarily cover genomic regions that are known to be functional, as is the case for gene expression or promoter arrays. Therefore tiling-arrays differ from these microarrays as they are not by definition designed to cover known or predicted genes in the genome. Moreover, the coverage of probes in unknown genomic regions has been useful for exploring the genome in an unbiased fashion. Examples of applications for tiling-arrays are: 1) protein-DNA-interaction by conducting chromatin immunoprecipitation (ChIP-on-chip) experiments[143], 2) epigenetic modifications by Methyl-DNA immunopreciptitation[56] (MeDIP-on-chip) or 3) identification of DNAse hypersensitive sites, which can be used to predict regulatory elements such as promoter regions, enhancers and silencers[144]. Although tiling-arrays are useful for genome wide studies, the coverage of the genome on the arrays depends on the species that is being studied. As an example, probes can cover the majority of a small genome such as for Arabidopsis [145] whereas probes will cover only contigs in a large genome, such as for human. Thus for larger genomes, as is the case for mouse or humans, the choice of the content depends on the questions one wishes to address using a particular tiling-array.

Each tiling-array produces quantitative signal intensity for each probe by the hybridization of labeled DNA. Normalized probe intensities are illustrated by the different peaks in Figure 1, where the colors indicate the probe signals at different chromosomes. Although single probe-hybridization with high signal intensity suggest strong hybridization, it

is not necessarily the result of specific hybridization of labeled DNA (illustrated by the probes above threshold 1 in Figure 1A and 1B). Multiple contiguous probes that show increased signal intensity upon hybridization across a particular genomic region are more likely to be the result of true hybridization in a biological experiment. These genomic regions are denoted as a putative region-of-interest (ROI). In order to find such ROI, a low threshold must be employed which may compromise the results by introducing false positives ROI (Figure 1A and 1B, threshold 2). To detect biological relevant ROI, probe intensity signals should be discriminated from non-specific signals. A challenge in the analysis of tiling-array data is the detection of true ROI, and to minimize the number of false positives. A straightforward approach is to choose a fixed number of consecutive probes above a certain threshold and indicate it as a ROI. Nevertheless, this definition of ROI may be inadequate because of the required number of consecutive probes and the optimal threshold may be difficult to establish. In addition, the probe-resolution varies across the genome, and across different tiling-array platforms.

Multiple methods have been developed to analyze tiling-array data which all serve one goal, i.e., the detection of true ROI and thereby discriminating positive probe intensity from the background. The developed methods differ in their statistical approaches: methods incorporate the Hypergeometric distribution[89], hidden Markov models[146-148], correlation structures[149], heuristics[150], mixture models[151], Bayesian modeling[152,153], wavelets[154], or by using other methodologies[55,74,155-160]. All methods have shown to be useful in filtering large data sets for candidate gene discovery. It is of importance to note, that biological experiments are always a necessity to validate particular findings.

Here we discuss the previously developed method, Hypergeometric Analysis of Tiling-arrays (HAT)[89], that uses the Hypergeometric distribution to assess the probability of a consecutive number of probes in a particular genomic region while controlling multiple testing (Family Wise Error: FWE). Furthermore, HAT uses multiple threshold cut-offs, it does not necessarily require experimental replicates, and can be normalized against reference files. It furthermore employs a single user defined parameter: the significance level alpha. Note that alpha is not used to determine the threshold cut-off using the data distribution (Figure 1B), instead it computes the probability to observe a specific number of probes for a particular genomic region (window) over multiple threshold cut-offs. Furthermore, specifying parameters such as fragment size may improve the detection of ROI, whereas parameters for gene mapping and sequence of interest are required for additional analysis (Figure 2). HAT is generically built and therefore independent of probe intensity distribution, probesets coverage and probesets resolution across the genome and tiling-array platform. It is successfully applied in multiple types of biological research questions, i.e., the detection of protein-DNA-interactions (ChIP-on-chip[89]), identification of genomic locations that are involved in viral integration and potentially harbor tumor suppressor genes (MeDIP-on-chip)[56], the identification of regions enriched for histone modifications such as, trimethylation of histone 3 at lysine 4 or lysine 27 (H3K4 me3, H3K27 me3)[89], and for the identification of anthocyanin-specific genes that flank enriched genomic DNA in black rice using 3'-TILLING 135 K *Oryza sativa* microarray[57]. Many detected ROI among these different studies were confirmed by quantitative polymerase chain reaction (qPCR).

Although tiling-arrays have been applied successfully for genome wide applications, high throughput sequencing of for instance chromatin immunoprecipitated DNA-fragments (ChIP-Seq), show genome wide associations in higher resolutions and will therefore be superior to chip technology. Even though ChIP-Seq is becoming the standard for genome wide applications, numerous high quality tiling-array data sets are publicly available at the the gene expression omnibus website (GEO: http://www.ncbi.nlm.nih.gov/geo/). These can be of value to address particular research questions raised by investigators and to which HAT may be very useful. Furthermore, although HAT was initially developed for the analysis of tiling-array data, the application is not limited to the studies discussed in this Chapter, but can be applied for the analysis of ChIP-Seq data as well.

Here we stepwise discuss how to apply HAT to analyze tiling-array data. As case examples we used two publicly available ChIP-on-chip data sets. In addition we discuss two types of analysis that frequently follow-upon the detection of ROI, namely motif and pathway analysis.

MATERIALS

We previously reported the successful usage of HAT on two novel data sets[89]. Here we demonstrate HAT on previously reported STAT4-chromatin immunoprecipitation (ChIP-on-chip) experiments (n=2), compared to controls (n=2). Secondly, we use HAT to analyze the DNA-binding capacity of a C-terminal mutant C/EBPα (n=2), compared to controls (ER) (n=2). Both data sets are available on the gene expression omnibus (GEO), GSE19321 and GSE16845 respectively. Data were generated using the Affymetrix GeneChip Mouse Promoter 1.0 Array. This chip generates 4.6 million perfect match probes over 28000 mouse promoter regions. Promoter regions cover 6kb upstream to 2.5kb downstream of 5' transcription start sites. Each probe has a size of 25 basepairs (bp). RAW probe intensity values are normalized by utilizing Model-based analysis of tiling-arrays for ChIP-on-chip (MAT) [74,88].

ANALYZING TILING-ARRAY DATA SETS

In this paragraph we demonstrate the usage of HAT for the identification of significant ROI and define the parameters for ChIP-on-chip experiments. Before starting the peak-detection algorithm (HAT), pre-knowledge about the experimental setup is highly recommended. The experimental protocol requires shearing of the DNA by using a sonication process which results in DNA-fragments of approximately 600 base pairs (bp). Subsequently, chromatin fragments are *immunoprecipitated* using antibodies directed to the protein-of-interest, known to interact with DNA. The consecutive probes can therefore cover up to 600bp after the hybridization process per fragment. This information can be used in the model for the detection of ROI. Note that significant ROI can be detected that are larger or smaller in width than 600bp. In addition, we set the significance level on 0.05.

The first ChIP-on-chip data set to which we applied HAT is a study that was previously reported and in which STAT4-mediated transcriptional regulatory networks in Th1 cell development were investigated[143]. STAT4 is a critical

component in the development of inflammatory adaptive immune responses. Although STAT4 was subject in various other studies[161,162], it was claimed that the genetic program, activated by STAT4 that results in an inflammatory cell type, is not well characterized. A ChIP-on-chip experiment was therefore conducted as previously reported[143]. Here, we analyzed both experimental replicates by choosing a fragment size of 600nt and α: 0.05, and detected n=2903 and n=3106 ROI. Moreover, 84% (n=2499) overlapped in both replicates compared to the controls (sized between 215bp-4543bp, median: 1002bp). It was previously demonstrated that the analysis method, GenPathway, identified 4669 genes that were seen in both replicates[143]. This list is subsequently filtered for genes with binding intensity > 4 and thereby resulted in 1540 genes. This indicates that using the unfiltered list, GenPathway detects almost twice the number of ROI when compared to HAT. To investigate the validity of the ROI that were detected by HAT, a motif enrichment analysis was conducted on the 2499 common ROI by using F-MATCH[163,164]. We hypothesize that the detected ROI should contain a STAT-binding site. We detected a total of 38 transcription factor binding sites of which the STAT-motifs were highly enriched (P<0.001). Moreover, 7 STAT-motifs were detected in the top 10 after ranking the TFBS on significance (Table 1). This suggests high specificity of the detected ROI. Note that the STAT-motif is also highly enriched in the genes detected by GenPathway[143]. Although both methods detected high enrichment for the STAT-motifs, the overlap of genes between both methods was 897 genes. In other words, 1211 genes were solely detected by HAT and not by GenPathway. To assess the validity of these ROI, we conducted a motif analysis for only those 1211 ROI and detected again high enrichment for the STAT-motifs, i.e., 6 STAT-TFBS are detected in the top ten ranked list (Table 2). We hypothesize that the 1211 genes may be present in the initial 4669 genes detected by GenPathway, but are excluded from the list as these did not comply the above mentioned criteria. This is supported by the notion that significantly lower probe intensity levels are observed (P<0.0001) in the 1211 ROI compared to the 897 ROI. Note that the probe intensity levels, of all the detected ROI, are significantly higher compared to the background. Unfortunately, we were not able to analyze the motifs among the exclusively detected genes by GenPathway, as the exact genomic positions of the ROI were not specified. These differences may occur due to alternatively defined gene mapping procedures Figure 3) and the differences in statistical methodologies. In conclusion, we identified another set of genes that were highly enriched for the STAT-motif.

The second ChIP-on-chip data set is used to study the DNA-binding capacity of a variant of CCAAT enhancer binding protein alpha (C/EBPα) that carries a C-terminal-mutation. C/EBPα is a transcription factor and master regulator of myeloid differentiation [165,166]. It is frequently mutated in patients with acute myeloid leukemia (AML) (5%-14%)[44]. Abnormalities in *CEBPA* may contribute to a block in differentiation of progenitor cells of granulocytes, which can result in leukemogenesis. Mutations in *CEBPA* are associated with a particular prognosis of patients with AML [44]. In AML patients, two types of *CEBPA* mutations are known to exist: mutations in the N-terminus and C-terminus. C-terminal mutations are found in the DNA-binding domain. Since the mutant protein can still interact with other proteins that may interact with DNA, we propose that mutant C/EBPα may indirectly interact with DNA. We wondered to which loci mutant C/EBPα might interact in an indirect manner. We created a similar C-terminal-mutation as found in one particular human AML patient[36], with an insertion of 6 amino acids in the C-terminal bZIP domain. We used it

in the ChIP-on-chip experiment to identify genes that may play a role in leukemogenesis. Promoter array hybridizations were conducted from a myeloid cell line model (32D), that expresses either beta-estradiol inducible C-terminal mutant C/EBPα (2 clones) or control-ER (2 clones). The question that we wished to address is whether mutated C/EBPα can bind to the DNA, thereby identifying the associated genes. Using a fragment size of 600bp and an alpha of 0.05, we detected in total n=89 and n=109 significant binding regions in the two clones with C-terminal mutant C/EBPα that was not seen in the controls Figure 4). The ROI are sized between 154 and 2481 nucleotides (median 717bp) and forty-eight were commonly detected in both clones.

We next searched for binding motifs among the detected ROI of the C-terminal mutant C/EBPα. Although it is known that the C-terminal mutant C/EBPα lacks binding capacity, we identified three enriched motifs namely, core-binding factor (CBF), ETS and ESE-1 (*P-value*<0.001). Core-binding factors have been shown to fulfil an important role in haematopoiesis[167] and ETS family members, such as ESE-1, fulfil an important role in several signal transduction pathways[168-170]. As expected, we did not find the consensus binding motif CEBP as we showed previously for wild-type C/EBPα using the same model system [89]. The detection of these three enriched motifs and the absence of the CEBP motif suggest that DNA-binding by mutant C/EBPα had occurred indirectly. We hypothesized that other factors may influence the DNA-binding capacity and therefore analyzed the 2kb upstream regions, from the transcriptional-start-site (TSS) of the detected genes (Figure 4). This resulted in the detection of 71 enriched TFBS with $P \leq 0.001$ and 1.5 times more frequently observed than in the reference set (fold-increase ≥ 1.5). As a reference set we selected 2kb upstream sequences (starting from the transcription start site) of 5000 randomly selected genes. The 2kb upstream sequences are gathered using the UCSC database (http://hgdownload.cse.ucsc.edu). The top 15 TFBSs are depicted in Table 3.

DETECTED REGIONS-OF-INTEREST CAN BE MAPPED TO GENES THAT ARE LOCATED IN CLOSE VICINITY

Although the goal is to detect ROI by using ChIP-on-chip tiling-arrays, it often requires additional analysis, such as pathway analysis, to test a particular hypothesis. This requires the mapping of ROI to genes. Each ROI can, theoretically, be mapped to four genes that are located on: 1. the positive strand and upstream, 2. the positive strand and downstream, 3. the negative strand and upstream, and 4. the negative strand and downstream (Figure 3). From these four genes, only one gene may be targeted (or two genes in a bi-directional promoter region). For promoter tiling-arrays, where only the promoter regions are present on chip, it is straightforward to map the detected ROI to the nearest located transcriptional-start-site (TSS) of a gene. To prevent incorrect gene mapping, due to differences in genomic locations of TSS between species and/or genomic-build (hg18, hg19 for human and mm8, mm9 for mus musculus, it is highly recommended to use the same species and genomic-build for both the gene mapping file as the one used in the normalization process. These gene mapping files can be downloaded from the UCSC: http://hgdownload.cse.ucsc.edu.

Manually curating each detected ROI to a particular gene is possible using the UCSC genome browser track (generated using HAT, Figure 2) but can be time consuming. Alternatively, by specifying the species and genome build in HAT, each ROI can automatically be mapped to the TSS of a gene in closest vicinity. We specified in both ChIP-on-chip experiments "mm8" because the experimental samples were derived from mus musculus and normalized with genomic-build 8. Because both analyzed data sets have been generated using promoter tiling-arrays, it allowed the mapping of the ROI to genes in close vicinity. For the STAT-study, the 2499 detected ROI were mapped to 2108 unique genes. For the C/EBPα-study, it resulted in the detection of 140 unique genes. These are graphically illustrated using a circos-plot[96] (Figure 4). Such graphical representation indicates the chromosomal location of the genes, and whether genes are commonly detected in the independent experiments using different clones.

MOTIF AND PATHWAY ANALYSIS ON THE DETECTED REGIONS-OF-INTEREST AND THEIR FLANKING GENES

Analysis on the detected ROI or the genes that are located in close vicinity of the ROI, is an important next step for hypothesis testing. Both motif and pathway analysis are therefore useful in tiling-array studies (Figure 2).

Motif analysis detects specific sequences involved directly in protein-DNA-binding interactions, or alternatively whether the promoter regions of the flanking genes include overrepresented sequences of transcription factors. These so called transcription factor binding sites (TFBS) may suggest that the protein-of-interest interacts synergistically with other proteins or is involved in the formation of protein-complexes. In general, two types of motif analysis exist: by using known TFBS that are derived from published collections (e.g., JASPAR or TRANSFAC databases). These databases should be used when seeking specific factors or structural classes. Secondly, dé-novo motif analysis can be used to analyze similarities among the sequences to produce a description for each pattern it discovers. F-MATCH[163,164] and MEME[112] are two algorithms which can be used for the detection of known TFBSs and/or dé-novo motifs. These methods are online accessible and require FASTA files as an input, which contain sequences of the ROI (generated by HAT).

Besides motif analysis, it can be useful to analyze the detected genes for enriched pathways. Pathway analysis is the process of identifying interactions and associated annotations[171]. For the detected flanking genes it may provide insight how genes are regulated and which processes, functions or networks were involved. Both commercial and noncommercial entities provide pathway analysis. A commercial tool is Ingenuity Pathway Analysis (Ingenuity® Systems, http://www.ingenuity.com, IPA 8.8). Networks in IPA are created using literature-based records that are maintained in the Ingenuity Pathway Knowledge Base. It computes a network-score for the overlap of the focus genes with a global molecular network. Alternatively, Gene Set Enrichment Analysis (GSEA)[97] provides both software and a collection of annotated gene sets (MSigDB: Molecular Signature Database) that can be used for the detection of pathways and/or gene sets (noncommercial). Depending on the research question, different gene sets can be used: 1. BioCarta pathways, describing the molecular relationships derived from active research areas, 2. KEGG pathways,

describing the molecular interactions and reaction networks, 3. Reactome pathways, manually curated and peer-reviewed pathways, 4. GO biological processes, gene sets describing the biological process ontology, 5. Transcription factor targets (TFT), gene sets contain genes that share a transcription factor binding site (TFBS), and 6. MicroRNA targets, Gene sets that contain genes that share a 3' UTR microRNA-binding motif.

NOTES

Different methodologies come to different results, what is the correct one to choose?

All previously described methods have been reported to validate some of the detected ROI as described in the first section, "*Introduction*". Nevertheless, different statistical methodologies lead to differences in the detected ROI. We hypothesize that various methodologies may results in similar detected ROI which are most likely the genomic regions that contain a contiguous number of probes with high probe intensity levels (the results of two methods are shown in section 3 "*Analyzing tiling-array data sets*"). In addition, the differences between detected ROI among various methodologies are likely the genomic regions with subtle changes in probe intensity levels. Note that some developed methodologies are designed for the analysis of one type of tiling-array application. Others may require various parameters to set before starting the analysis, e.g., by defining the ROI using the maximum and/or minimum number of probes in a genomic region, maximum gap size between two probes and threshold. Changing one of the parameters will affect the final results. It is therefore always recommended to perform additional analysis after the detection of ROI to ensure confidence about the gained results. We demonstrated this in section 3 "*Analyzing tiling-array data sets*", where we detected 1211 ROI that were exclusively found for HAT. A motif analysis showed significant enrichment for the STAT-consensus binding site. Such findings may help deciding which method to use. It is important to note that in the end laboratory experiments are indispensable to demonstrate the biological significance of particular that ROI, identified by means of tiling-array analysis.

How to continue if no significant regions-of-interest are detected?

The analysis of tiling-array data (section 3 *"Analyzing tiling-array data sets")* can result in the absence of significantly enriched ROI. This indicates that probe intensity values, by the hybridization of DNA-fragments on chip, showed no significant differences compared to the background data-file. In case the hybridization process on chip is successfully performed (i.e., DNA-fragments are immunoprecipitated) and the background data-file is correctly provided into the model, it still may result in the absence of significantly enriched ROI. Note that analysing experimental data-files without the usage or incorrect usage of a background data-file can lead to the absence of significantly enriched ROI or the detection of false positive ROI. If no significantly enriched ROI are detected, it should be considered that no DNA-binding did take place and therefore no ROI were detected. Alternatively, one could decide to increase the significance level alpha and rerun the analysis. Note that the false positive rate increases by using alpha>0.05. It is therefore highly recommended to validate the ROI by qPCR. As an example, it is demonstrated that a MeDIP-on-chip

experiment resulted in the detection of 15 ROI[56]. These are detected without using a background[56]. Although there was supporting evidence that all 15 ROI may be valid (additional analysis showed that all ROI contained a nearby restriction site), only eight viral integration sites could be validated by directed PCR followed by Sanger sequencing[56]. The remaining seven ROI may therefore be the result of technical variation which may have been prevented by using a background file.

How to determine the best flanking gene for a detected regions-of-interest?

The usage of tiling-array data does not provide information regarding the strand (positive or negative) or genes affected by the putative promoter. It only indicates the probe intensity values and their genomic positions. If a particular genomic region is marked as a potential ROI, the responsible immunoprecipitated DNA-fragment is suggested to show binding, e.g., via an immunoprecipitated transcription factor, that could bind to the DNA strand. The ROI is then linked to the gene in close vicinity (Figure 3). The use of an UCSC-browser track may help manually curating the ROI to a gene. Alternatively, it requires biological experiments to validate whether the binding had an effect on the regulation of a gene. Note that promoter tiling-arrays (as described in section 2 *"materials"*) only contain probes of which the genomic locations are in the promoter regions of genes and therefore simplifies the gene mapping procedure.

How to run HAT with RAW cell files?

HAT is built generically to analyze different applications and platforms of tiling-array data (as described in section "*Introduction*"). On the contrary, normalization may differ between different applications and platforms of tiling-arrays, e.g., one-color arrays of Affymetrix versus two-color arrays of Nimblegen. Including a normalization step into the model would therefore limit the model to one type of tiling-array. RAW cell files need to be normalized based on the type of tiling-array[88], and then used as an input into the model (Figure 2).

How to prevent “out-of-memory” problems when analyzing tiling-array data?

When using HAT, it is recommended to use at least 4GB of RAM memory and Windows-64bits version or UNIX-based system. The methodology is tested on tiling-array data containing 4.6 million perfect match probes, and developed in such a way that it is analyzes per chromosome which reduces high memory loads. Nevertheless, when memory problems occur, it is recommended to kill unused running processes when running HAT. In a Windows environment this can be done in the “task manager”; find the "*Run*" window in the start-menu en type "*taskmgr*" and then press “*Ok*” or press the <ENTER>-key.

How to install HAT in Windows or an UNIX environment?

The installation of HAT requires an x86-64 Windows or UNIX-based system and 4GB memory or more is highly recommended. Both platforms require the installation of MATLAB or the freely available MATLAB Compiler Runtime

(MCR) which is a standalone set of shared libraries that enable full functioning of HAT. Documentation regarding the installation procedure can be found on: http://www.erasmusmc.nl/hematologie/ or http://hema13.erasmusmc.nl/index.php/HATSEQ

ACKNOWLEDGEMENT

We thank Claudia A. Erpelinck-Verschueren for technical assistance in the preparation of the C/EBPα C-terminal mutant samples.

FIGURE LEGENDS

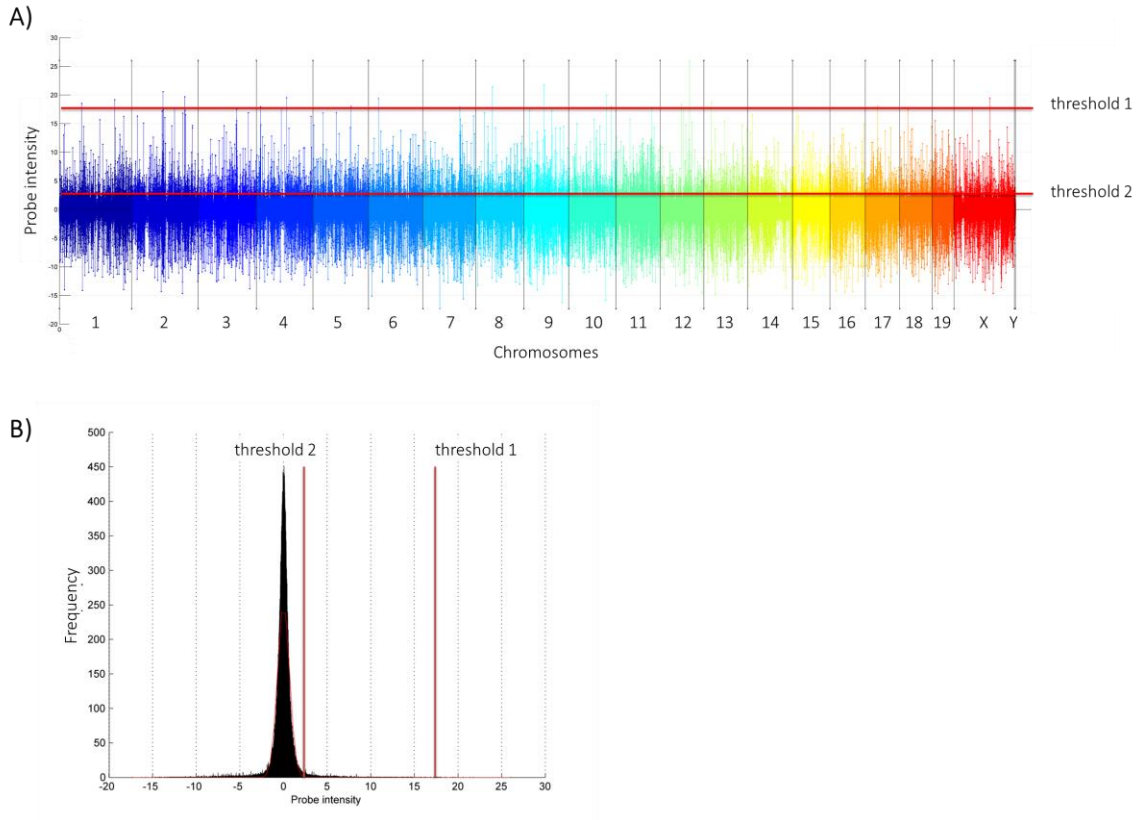


Figure 1. Graphical representation of probe intensities in a ChIP-on-chip tiling-array experiment. (A) Normalized probe intensity of 4.6 million probes among 22 chromosomes. Colours illustrate the different chromosomes whereas the length of a lollipop represents the probe intensity. (B) Distribution of the probe intensity values. The probe intensity values are normalized against a reference file. Threshold 1 indicates a high threshold cut-off whereas threshold 2 indicates a low intensity cut-off. HAT uses many different threshold cut-offs to determine significantly enriched ROI.

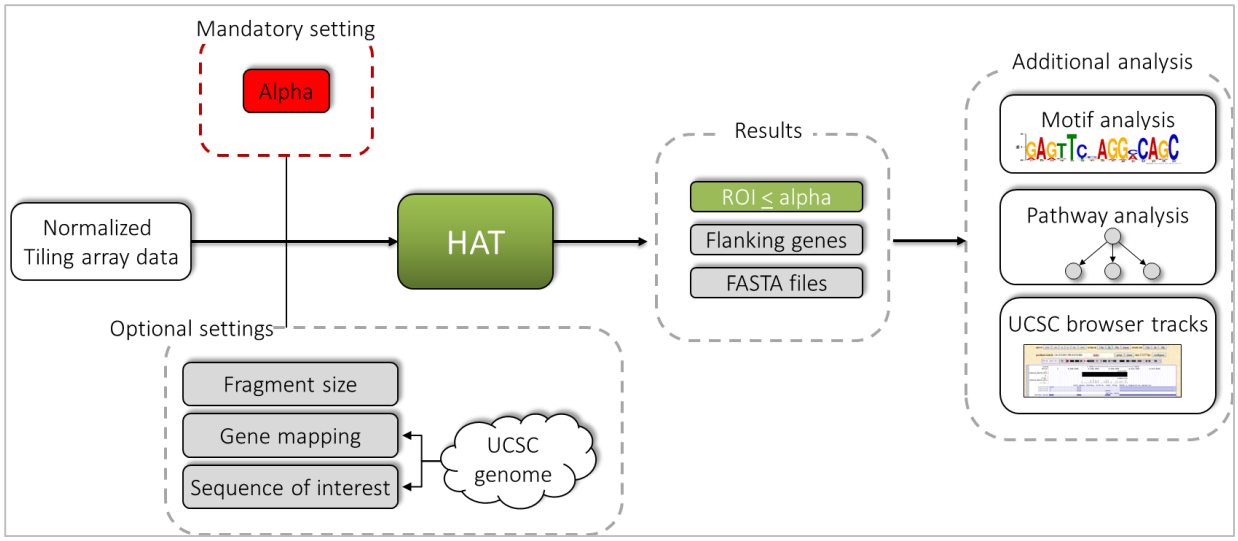


Figure 2. Schematic overview of tiling-array data analysis. Stepwise illustration of normalized tiling-array data towards the detection of significantly enriched ROI, the flanking genes, sequences files (FASTA), motif analysis, pathway analysis and the UCSC-browser.

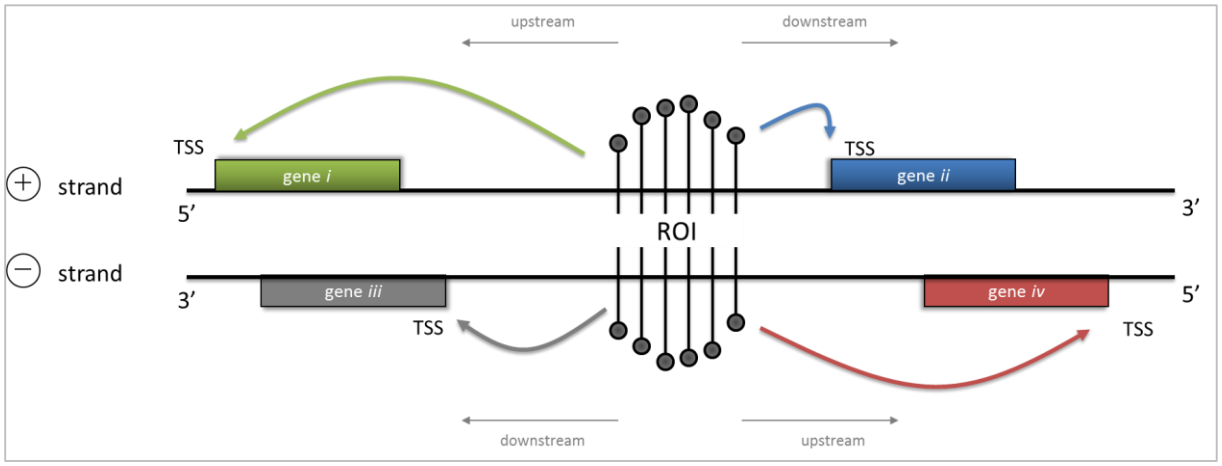


Figure 3. Mapping of detected regions-of-interest to genes located in close vicinity. A single ROI is illustrated with four neighbouring genes: two on the positive-strand (upstream and downstream) and two on the negative-strand (upstream and downstream). Mapping of ROI to genes is crucial for additional analysis (e.g., pathway analysis). Abbreviations, ROI: Region-of-interest, TSS: Transcriptional-start-site, 3': Three prime UTR, 5': Five prime UTR.

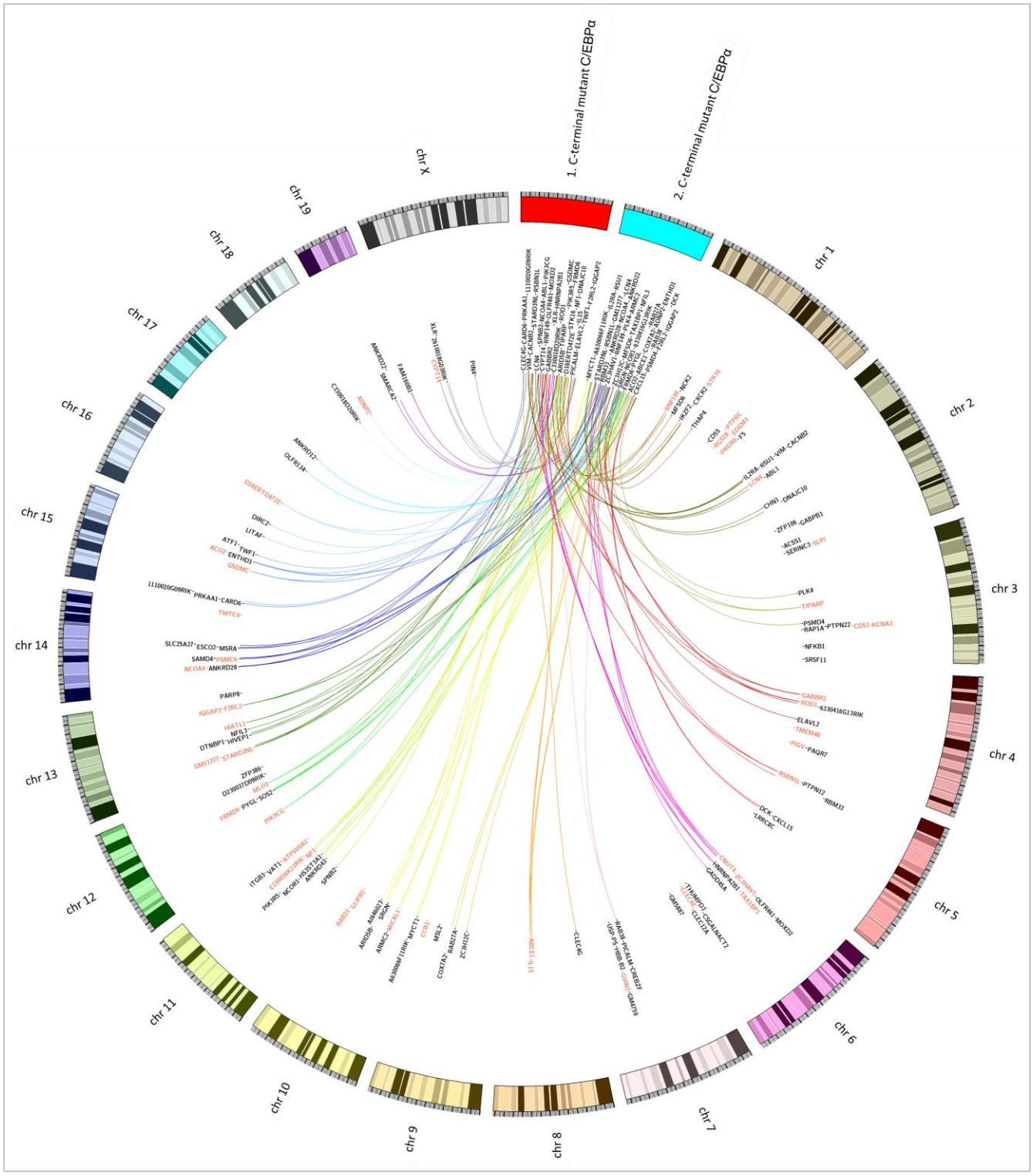


Figure 4. Graphical representation of the genes that are bound by the C-terminal mutant C/EBPα. One hundred and forty mapped genes from the detected ROI of the C-terminal mutant C/EBPα experiments are illustrated. Candidate genes in experiment 1 are indicated by the red box whereas the candidate genes from experiment 2 are indicated by

the blue box. Forty-six genes (mapped from 48 ROI) that overlap between experiment 1 and 2 are indicated with a red text-color. Line-colors are colored similar as the chromosomes which are numbered from 1-19 and X, and show the relative location of the genes using mouse genome build 8 (mm8).

Transcription Factor	Recognized factors	Fold-increase	*P*-value
V$STAT1_01	STAT1, STAT1alpha, STAT1beta	8.059	5.53E-29
V$STAT5B_01	STAT5A, STAT5B	4.092	1.05E-26
V$STAT1_05	STAT1	5.611	2.15E-26
V$STAT_01	STAT1, STAT1alpha, STAT1beta, STAT2, STAT3, STAT3-isoform1, STAT4, STAT5A, STAT5B, STAT6	3.542	5.45E-22
V$STAT3_01	STAT3, STAT3-isoform1	5.963	1.34E-19
V$STAT1STAT1_Q3	CBF3, STAT1:STAT1, ehf	4.046	1.11E-18
V$IRF_Q6	IRF-10, IRF-2, IRF-3, IRF-4, IRF-5, IRF-6, IRF-7, IRF-7A, IRF-7B, IRF-7H, IRF-8, IRF4-1, irf1	3.603	1.08E-11
V$AP1_Q6_01	AP-1, FOSB, FosB, Fra-1, Fra-2, JunB, JunB:Fra-1, JunB:Fra-2, JunD, JunD:Fra-2, JunD:deltaFosB, c-Fos, c-Jun, c-Jun:FosB, c-Jun:JunD, c-Jun:c-Fos, deltaFosB	2.703	2.04E-11
V$STAT5A_01	STAT5A	4.080	3.81E-11
V$BACH1_01	Bach1, Bach1t	3.121	3.35E-09

Table 1. Motif enrichment analysis on the detected regions-of-interest in the STAT4 experiment. The top 10 enriched TFBS among the detected binding regions using HAT for the STAT4-study (ChIP-on-chip). A TFBS is called when the position weight matrices (PWM) are enriched at $P \leq 0.001$. Recognized factors: the transcription factors that are recognized by the TFBS. Fold-increase: the frequency that a TFBS is detected among the binding regions compared to the reference set (5000 randomly chosen genes).

Transcription Factor	Recognized factors	Fold-increase	*P*-value
V$STAT1_01	STAT1, STAT1alpha, STAT1beta, STAT2, STAT3, STAT3-isoform1, STAT4, STAT5A, STAT5B, STAT6	7.235	4.96E-16
V$STAT3_01	STAT3, STAT3-isoform1	5.862	7.36E-12
V$STAT5B_01	STAT5A, STAT5B	3.633	2.16E-11
V$STAT1_05	STAT1	4.658	2.45E-09
V$STAT_01	STAT1, STAT1alpha, STAT1beta	3.405	4.05E-09
V$GADP_01	GABP	4.416	1.88E-08
V$SAP1A_01	SAP-1a	4.106	5.32E-08
V$STAT1STAT1_Q3	CBF3, STAT1:STAT1, ehf	3.507	5.26E-07
V$ELK1_02	Elk-1, Elk1-isoform1	4.025	7.68E-07
V$CETS1P54_01	Ets-1, Ets-1 deltaVII, c-Ets-1, c-Ets-1 54, c-Ets-1A, c-Ets-1B	3.981	8.78E-07

Table 2. Motif enrichment analysis on the detected regions-of-interest that are exclusively detected using HAT in the STAT4 experiment. The top 10 enriched TFBS among the exclusively detected binding regions of HAT for the STAT4-study (ChIP-on-chip). A TFBS is called when the position weight matrices (PWM) is enriched at $P \leq 0.001$. Recognized factors: the transcription factors that are recognized by the TFBS. Fold-increase: the frequency that a TFBS is detected among the binding regions compared to the reference set (5000 randomly chosen genes).

Transcription Factor	Recognized factors	Fold-increase	*P*-value
V$POU3F2_02	POU3F2, POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b)	1.948	6.81E-12
V$CDP_02	CDP, CDP-isoform1, CDP2	2.072	9.86E-09
V$FOXP3_Q4	FOXP3	4.753	5.98E-08
V$OCT1_01	Oct-1, POU2F1, POU2F1a	1.713	6.24E-08
V$IPF1_Q6	PDX1, ipf1	1.605	1.37E-07
V$CLOX_01	Cutl	1.756	8.19E-07
V$SATB1_01	CBF-C	1.587	8.48E-07
V$OTX_Q1	Otx1, Otx2	1.903	9.09E-07
V$HMGIY_Q3	HMGI-C, HMGIY, HMGIY-isoform1, HMGIY-isoform2	1.546	1.72E-06
V$FOXO1_01	FOXO1A	1.714	1.93E-06
V$DMRT4_01	DMRT4	1.581	2.29E-06
V$NFAT_Q6	NF-AT, NF-AT1, NF-AT1C, NF-AT2, NF-AT3, NF-AT4, NFAT1, NFAT1-isoformD	3.009	2.74E-06
V$TEF_Q6	TEF-xbb1, Thyrotroph embryonic factor, Thyrotroph embryonic factor-isoform1, Thyrotroph embryonic factor-isoform2, Thyrotroph embryonic factor-isoform3, VBP	1.955	5.14E-06
V$SRF_C	SRF, SRF-I, SRF-L, SRF-M, SRF-S	2.111	5.91E-06
V$CEBPGAMMA_Q6	C/EBPgamma	1.828	6.50E-06

Table 3. Motif enrichment analysis on the 2kb upstream regions-of-interest of the C-terminal mutant C/EBPα target genes. The top 15 enriched TFBS among the 2kb upstream genes of the C-terminal mutant ROI. A TFBS is called when the position weight matrices (PWM) are enriched at $P \leq 0.001$ and with fold-increase>1.5. Recognized factors: the transcription factors that are recognized by the TFBS. Fold-increase: the frequency that a TFBS is detected among the binding regions compared to the reference set (5000 randomly chosen genes).

CHAPTER 6

A Repressor Function of C/EBPα is Indicated by using Combined Gene Expression Profiling in AML and Chromatin Immunoprecipitation Data

In preparation

A Repressor Function of C/EBPα is Indicated by using Combined Gene Expression Profiling in AML and Chromatin Immunoprecipitation Data

Erdogan Taskesen, Bas J. Wouters, Roberto Avellino, Meritxell AlberichJorda, Daniel G. Tenen, Jeroen de Ridder, Peter J.M. Valk, Claudia A. Erpelinck-Verschueren, Marcel J.T. Reinders and Ruud Delwel

ABSTRACT

C/EBPα is a transcription regulator that is essential for normal neutrophil development. We hypothesize that methylation and consequent silencing of the gene encoding C/EBPα is an abnormal event that occurs in a subset of human acute leukemias. This recently identified *CEBPA*^{silenced} group subtype has a unique epigenetic feature, i.e. silencing of the gene that encodes CCAAT-enhancer binding protein alpha (C/EBPα) by DNA hypermethylation. The leukemic blast cells of these patients express myeloid as well as T-lymphoid markers. Moreover, gene expression and DNA-methylation profiling stratifies these leukemias in between Acute Myeloid Leukemia (AML) and T-lymphoid Acute Lymphoblastic Leukemia (T-ALL). We carried out gene expression profiling of the *CEBPA*^{silenced} group and identified the genes that were differentially expressed in this AML subtype compared to other AMLs and normal marrow blast cells. To assess whether these genes are C/EBPα targets and whether expression has been altered as the result of *CEBPA* loss of expression, we transduced an estrogen-inducible C/EBPα construct in 32D cells and carried out ChIP-on-chip using ER specific antibodies. We detected 529 C/EBPα target genes that were subsequently overlaid with the differentially expressed genes in the *CEBPA*^{silenced} group. This resulted in 49 overlapping genes (P=1x10^{-7}) that are indicative as putative direct targets. We hypothesized that the downregulated genes, that overlap with the C/EBPα target genes (n=25, P=1.2x10^{-3}) represent targets that are normally activated by wild-type C/EBPα. The upregulated genes, that overlap with the C/EBPα target genes (n=24, P=1.6x10^{-5}) are assumed to be repressed in wild-type C/EBPα expressing cells and activated when *CEBPA* is silenced. A selection of the latter group of genes appeared to be upregulated in hematopoietic stem cells of *Cebpa* knock-out mice as well, emphasizing a putative repressor function of this transcription factor for certain genes. We hypothesize that *CEBPA* silencing is a transforming event that allows the expression of lymphoid genes to take place in a subset of leukemias with myeloid/T-lymphoid features. These genes are predicted to be repressed by direct promoter interaction with C/EBPα in normal myeloid progenitors and in other AMLs.

INTRODUCTION

Myeloid committed progenitors are under tight control of combinations of transcription factors that modulate gene expression in order to maintain a balance between differentiation and proliferation. Transcription factor

CCAAT/enhancer binding protein alpha (C/EBPα) is one of the master regulators of myeloid differentiation[165,166]. C/EBPα mRNA encodes a 42kDa (p42) protein, and a shorter isoform of 30kDa (p30)[53,172-174]. Both isoforms share a highly conserved C-Terminal domain that contains a basic region required for DNA-binding and a leucine zipper (bZIP) essential for homo or heterodimerization. The N-terminal domain is less conserved and consists of two transcription activation domains (TADs) that contribute to cell growth inhibition. Both isoforms appear to be expressed at a constant ratio which is possibly required to maintain the function of C/EBPα and regulate normal differentiation. Deregulation of *CEBPA* expression causes an imbalance in myeloid differentiation[175]. In *Cebpa* knock-out mice, neutrophil development is disrupted and consequently only myeloblasts are detectable in the marrow of these mice[175]. The functional importance of *CEBPA* in myelopoeisis is further supported by its deregulation in various subsets of AML patients[18]. In AML with recurrent translocations t(8;21), expressing the translocation specific *AML1-ETO* fusion gene, *CEBPA* has been reported to be downregulated by the oncogenic fusion protein[176]. Mutations in *CEBPA,* the gene encoding C/EBPα, have been demonstrated in multiple studies[52,53,177]. These mutations have been found in the N-terminus or in the C-terminus of the protein, and can either occur in a monoallelic or biallelic fashion[178,179]. As a consequence, these mutations impair differentiation of hematopoietic progenitors[40,180].

A subset of AML patients, previously characterized as *CEBPA*^{silenced} AMLs, lack *CEBPA* expression due to an aberrant DNA hypermethylation pattern. This subset of patients have a distinct phenotype when compared to other AMLs; their leukemic cells express myeloid surface antigens CD13 and CD33 as well as T-lymphoid markers. The most consistently expressed surface protein is *CD7*, while other T-cell related genes are also expressed such as *LCK*, *TRIB2*, *CD3D*, *CD3G*, *TRD@* and *NOTCH1*[61]. We recently demonstrated a similar reverse correlation between *CEBPA* and lymphoid genes in myeloblasts from *Cebpa* knock-out mice[175 61]. These findings made us hypothesize that there is a negative regulatory control of a set of genes by C/EBPα and that these genes, of which many are T-cell related, are transcriptionally activated in case the expression of *CEBPA* is turned down. The data that we provide in this study are in support of a hypothesis of a direct negative regulatory control of a set of genes by C/EBPα.

MATERIAL AND METHODS

DATA

Two Large scale datasets were used in this study: ChIP-on-chip data using Affymetrix GeneChip Mouse Promoter 1.0 array, derived from an inducible *CEBPA* expressing myeloid cell line model of three wild-type C/EBPα-ER clones and two control clones (C/EBPα-mutant-ER) and three control clones expressing a construct that only contains ER. Data is available at the NCBI Gene Expression Omnibus (GEO), accession number GSE19321. Secondly, genome wide gene expression profiles were generated for 506 de novo AMLs and 11 normal CD34^{+}, using Affymetrix U133 Plus 2 microarray (Santa Clara, CA, USA) previously[18]. GEP data are available at the Gene Expression Omnibus (National Center for Biotechnology Information; accession number GSE14468 (HOVON-SAKK cohort). Sample processing and quality control were carried out as described previously[18,181]. Normalization of raw data was processed with Affymetrix

Microarray Suite 5 (MAS5) to target intensity values at 100. Intensity values were $\log_2$ transformed and mean centered per probeset.

Chromatin immunoprecipitation on DNA promoter microarrays

Chromatin immunoprecipitation (ChIP) was carried out according to a protocol from Affymetrix (Santa Clara, CA, USA). Genome wide discovery of C/EBPα targets was conducted using ChIP from a beta-estradiol induced C/EBPα in a myeloid cell line model (32D) followed by promoter array hybridizations (for details see supplementary material, section: C/EBPα-ER cells differentiate upon E2 treatment while C/EBPα-mutant-ER cells show impaired differentiation). This chip contains 4.6 million perfect match probes over 28000 mouse promoter-regions. Promoter-regions have 10Kb coverage for each promoter-region and each probe has a size of 25 nucleotides. Clones express either beta-estradiol inducible C/EBPα-ER or beta-estradiol inducible C/EBPα-mutant-ER. The mutant has an insertion of 6 amino acids in the bZIP domain and previously found in a human AML patient[36] (for details see supplementary material, section: Plasmids). It showed less pronounced inhibition of proliferation upon treatment with E_2 in the presence of IL3. In the presence of G-CSF, C/EBPα-mutant-ER cells also demonstrated delayed differentiation with the suggestion of a partial block. Chromatin immunoprecipitations were carried out using an antibody directed against ER in the beta-estradiol treated cells (E_2 for 4 hours) and the DNA obtained from these cells, after immunoprecipitation, was hybridized to Affymetrix promoter chips.

METHODS

Aberrant genes that are specific for the *CEBPA*^{silenced} group in human AML are derived by comparing the gene expression data of the *CEBPA*^{silenced} patients (n=10) against CD34$^+$ samples (n=11) and against the remaining AML group (n=496) using a three-way ANOVA and a post-hoc test. The post-hoc test is based on the tukey-kramer method which is used to select genes that significantly differed between: *1) CEBPA*^{silenced} group versus CD34$^+$ group and, *2) CEBPA*^{silenced} group versus the other AMLs. Genes are considered to be differentially expressed when mRNA levels differed with $P \leq 0.05$ after multiple testing (using the family wise error rate, FWER).

Binding of C/EBPα in 32D cells was determined by utilizing Hypergeometric Analysis of Tiling-arrays (HAT[182]) after normalization of the raw probe intensity data[74]. As a result of this normalization, probe intensity values follow a normal distribution with a negative mean. Probe intensities that may be the result of hybridization of labeled DNA on the chip, have values greater than zero and are processed to determine candidate C/EBPα binding regions. A binding event was called when fragments are enriched with significance level ≤ 0.05 and, maximum fragment size of 600bp. This fragment size correlates with the average sonicated fragment sizes, being 600bp. Furthermore, we considered only genes from which the enriched binding region to the transcriptional-start-site (TSS) was located within 2Kb and, binding of C/EBPα in the promoter-region was detected in two or more clones. Gene-symbols are annotated using HAT[182] and mapped using NCBI murine Genome Build 36 (February 2006). This resulted in the identification of 529

unique genes, by comparing C/EBPα-ER clones (n=3) versus C/EBPα-mutant-ER clones (n=2). It is expected that the mutations inserted in the bZIP domain of C/EBPα-mutant-ER clones disrupt the physical binding of C/EBPα with DNA motif sequences of downstream target genes. More detailed information about the detected binding regions and enriched motifs can be found in supplementary material (candidate C/EBPα targets: ChIP-on-chip promoter arrays).

Transcription Factor Binding Site analysis - For the identification of transcription factor binding sites (TFBSs), we used F-MATCH[163] to scan the 2Kb upstream regions, from the TSS, for occurrence of one or more of the 656 highly specific position weight matrices (high-PWMs) gathered from the TRANSFAC Pro database. A TFBS was called when its PWM is enriched at $P \leq 0.001$. As a background set we selected 2Kb upstream sequences (starting from the transcription start site) of 5000 randomly selected genes that are not associated with C/EBPα in terms of binding, using the 32D model-system or identified as differential expressed for the C/EBPα^{silenced} group. The 2Kb upstream sequences for each gene were retrieved from the UCSC database (http://hgdownload.cse.ucsc.edu/goldenPath/).

RESULTS

Genes involved in T-cell development are differentially upregulated in *CEBPA*^{silenced} leukemias.

To identify genes differentially expressed in *CEBPA*^{silenced} AMLs (n=10), we compared their gene expression profiles with those of CD34^{+} normal bone marrow (n=11) and of the other AML cases[18] (n=496). Six hundred eighty-nine differentially expressed genes were identified in *CEBPA*^{silenced} leukemias, of which 286 (blue circle in Figure 1A) were upregulated and 403 downregulated (green circle in Figure 1A). Twenty pathways were significantly represented by Ingenuity pathway analysis (IPA) using both up and downregulated genes (Figure 1B and Table S1). Genes involved in T-cell development appeared to be highly enriched. Notably, the genes upregulated in *CEBPA*^{silenced} leukemias were more frequently annotated with T-cell functions than the downregulated genes. Therefore we carried out pathway analysis of the 286 upregulated or 403 downregulated genes separately and identified 42 and 80 unique pathways ($P \leq 0.001$) respectively (Table S1). We indeed observed associations with (T-) lymphocytes for the enriched pathways specific for the upregulated genes in *CEBPA*^{silenced} leukemias (Figure S1A and Table S1). Pathways specific for the downregulated genes (Figure S1B and Table S1) included myeloid associations, such as pathways that are involved with macrophage, neutrophil or dendritic cell development. Besides analysing the pathways associated with molecular, cellular and developmental functions we also analysed the canonical pathways and detected a strong relation with T-cell development/function for the upregulated genes, in contrast to the downregulated genes in leukemias with *CEBPA* being switched off (Figure S2 and Table S2). These data suggest a relation between the absence of *CEBPA* and activation of T-cell related genes.

We next investigated the presence of transcription factor binding sites (TFBSs) in the promoter regions (2Kb upstream region of the TSS) for the 286 upregulated and 403 downregulated genes. We detected 13 and 4 unique significantly enriched TFBSs with fold-increase $\geq$ 1.5 respectively (Table S3). Among the upregulated or downregulated genes we

did not detect C/EBPα consensus binding sites (CEBP). We predict that only a fraction of the differentially expressed genes are true targets of C/EBPα, which would explain why no consensus site was found.

C/EBPα requires its DNA-binding domain to induce myeloid differentiation.

Abnormal myeloid maturation in the bone marrow of *Cebpa* knock-out mice defines the critical function of *CEBPA* as a potent inducer of myeloid differentiation. Upon the loss of *CEBPA* expression in *CEBPA*^{silenced} AML, it is hypothesized that a subset of genes involved in myeloid lineage commitment and differentiation are aberrantly regulated and differentially expressed. Briefly, cells were retroviral transduced with murine C/EBPα coding sequence fused to the ligand-binding domain (LBD) of the human estrogen receptor alpha (ERa) (C/EBPα-ER), acting as an inducible system upon estradiol (E_2) exposure. The LBD of ERa engages E2, become activated and relocate to the nucleus as a CEBPA-ER fusion protein. Once C/EBPα is in the nucleus, it accesses regulatory regions of downstream target genes that are responsible for myeloid differentiation.

E2-induced differentiation was observed when the cells were cultured in the presence of interleukin 3 (IL-3) (Figure S3A and B)) or G-CSF[183] (Figure S3C and D) showing that C/EBPα is a potent inducer of differentiation even at proliferative conditions when cells were cultured in the presence of IL-3. A construct expressing the LBD of ERa (ER) only (without C/EBPα) was introduced into 32D cells that did not differentiate upon E_2 exposure when cultured in the presence of IL3 (Figure S3B). In the presence of G-CSF these cells showed the expected neutrophil development (Figure S3D). To study requirement of DNA-binding of C/EBPα-ER protein, a mutant was constructed, which carried an insertion of 6 amino acids in the bZIP domain of C/EBPα (C/EBPα-mutant-ER). This mutant was identified in a human AML patient and was predicted to lack DNA-binding activity[27]. Morphological analysis revealed that differentiation of the C/EBPα-mutant-ER 32D cells was absent when stimulated with IL3 plus E_2 (Figure S3A and B). Strongly delayed differentiation was observed when cells were stimulated with G-CSF plus E2 (Figure S3C and D). We hypothesized that defective differentiation by mutant *CEBPA* was caused by a lack of ability to bind DNA and to regulate the expression of genes critical for myeloid differentiation. We postulate that these genes can be identified using chromatin immunoprecipitation (ChIP) and by comparing the binding of C/EBPα-ER versus C/EBPα-mutant-ER in 32D cells.

Discovery of C/EBPα interacting loci in C/EBPα-ER 32D cells.

Anti-ER ChIP was carried out for C/EBPα-ER and C/EBPα-mutant-ER 32D clones followed by Affymetrix promoter chip hybridization. The immunoprecipitated fragments were analyzed using HAT which resulted in the detection of C/EBPα-ER binding regions, i.e. loci that interacted with C/EBPα-ER but not with mutant C/EBPα-ER (Table S4) within 529 unique gene promoters. Among those previously described C/EBPα targets were present, as an example, the binding of C/EBPα to the *Il6ra*[82] promoter (Figure 2A). These results were validated using a quantitative ChIP-Q-PCR experiment for *Il6ra* (Figure 2B).

Transcription factor binding site analysis revealed that only the CEBP consensus binding site was significantly enriched (Table S5, $P<1x10^{-28}$) in the detected binding regions within the 529 promoters. The detected binding sites were uniformly distributed within the upstream sequences (Figure S4). Using this approach we identified several known but also many novel putative target genes of C/EBPα (Table S4).

Differentially expressed genes in *CEBPA*^{silenced} leukemias are enriched for C/EBPα target genes.

We next addressed the question whether among the 689 differentially expressed genes identified in *CEBPA*^{silenced} leukemias, direct C/EBPα targets were present that were identified by ChIP-on-chip using the C/EBPα-ER 32D model. We overlaid the 529 C/EBPα targets found in the 32D model-system (yellow circle in Figure 1A) with the 689 differentially expressed genes identified in *CEBPA*^{silenced} leukemias (green circle and blue circles in Figure 1A) and found an overlap of 49 genes (Figure 1A). The chance that two gene sets of this size show an overlap of 49 genes is $P=1x10^{-7}$. Of these 49 genes, 25 were downregulated and 24 upregulated in *CEBPA*^{silenced} leukemias ($P=1.2x10^{-3}$ and $P=1.6x10^{-5}$ respectively, Table 1). A selection of these genes are depicted in the heat map in Figure 3. These data suggest that upon binding to a promoter, C/EBPα may affect the expression of a putative target gene in different ways, i.e. it may act as a transcriptional activator or it may function as a repressor of transcription.

The putative C/EBPα target genes upregulated in *CEBPA*^{silenced} leukemias are relevant for T-cell development.

We hypothesize that among the 24 genes that are upregulated in *CEBPA*^{silenced} leukemias and bound by C/EBPα, are under normal conditions repressed in myeloid cells upon C/EBPα interaction. We already demonstrated that the 286 upregulated genes in *CEBPA*^{silenced} AMLs were enriched for T-cell associated genes. Pathway analysis for these 24 upregulated genes in *CEBPA*^{silenced} leukemias showed enrichment for pathways involved in T-cell development (Figure 1C and Table S6). The networks are illustrated in Figure S5A-C. This observation suggests that the transcription factor C/EBPα may act as a repressor of genes such as *BCL2*, *CCR9, B3GNT2, CD47, CASP1* or *MAP2K4*.

Twenty-five C/EBPα genes bound in the 32D model were absent in *CEBPA*^{silenced} human AML but transcribed in other myeloid leukemias. Apart from *TOB1,* these genes appeared not to be associated with T-cell related functions (Figure 1C, Table S6 and Figure S5D-E). In fact, *TOB1* is a gene that encodes an inhibitor of transcription of cytokines and cyclins and represses T-cell proliferation.

T-cell related genes are upregulated in *Cebpa* knock-out murine bone marrow progenitor cells

To study whether the regulation of the differentially expressed genes required C/EBPα in normal myeloid progenitors, a *Cebpa*-knock-out (KO) mouse model-system was applied. We purified short-term (ST, n=2) and long-term (LT, n=2) hematopoietic stem cell (HSC) from wild-type as well as *Cebpa*-KO bone marrow as decribed previously[184], isolated mRNA and carried out microarray hybridizations. We selectively studied the 49 genes (Table 1) and compared the mRNA-expression of those genes in ST-HSCs and LT-HSCs. We overlaid the 24 upregulated genes in *CEBPA*^{silenced} leukemias (Table 1) with the genes that showed a fold-increase of ≥ 1 in separately the KO-LT-HSC and KO-ST-HSC and

detected an overlap of 6 genes, including *CCR9* and *CEBPG* (Figure S6A-B). These genes are strongly upregulated in the absence of wild-type *CEBPA*, indicating a putative role for C/EBPα as a repressor in HSCs. For the 25 downregulated genes in *CEBPA*^{silenced} leukemias (Table 1), we detected downregulation for 11 genes in the *Cebpa*-KO marrow HSCs, among them *TOB1* (Figure S6C-D).

C/EBPα acts synergistically with other transcription factors for T-cell repression

To reveal transcriptional modules that contribute to the activation of genes, we analyzed the promoter regions of the 49 genes. The C/EBPα consensus binding sites were detected in the promoter regions of 20/24 and 16/25 upregulated and downregulated genes respectively. The promoter regions of the upregulated genes (n=24) revealed 27 significantly enriched TFBSs for 23 families. Besides the C/EBPα consensus site and sites that are described in literature to interact synergistically with wild-type C/EBPα (PU.1[185,186], SP1[187] and E2F[188,189]), we also detected 20 other TFBSs which are listed in Table S7. Interestingly, for E2F it has been reported that it interacts with *CEBPA*^{wt} and is involved in the downregulation of genes[188,189]. We furthermore detected E4BP4[190] for which is known to be involved in the regulation of T-cell interleukins. Transcription factor (TF) interactions are a crucial aspect of the gene regulatory system[191]. Thus the T-cell repressor function of C/EBPα may be performed by synergistically interacting with these transcription factors[185,186,188,189,192-194]..

DISCUSSION

In this study, we investigated the DNA-interaction of exogenous C/EBPα on target loci using CHIP coupled with genome wide promoter arrays. We determined C/EBPα interaction in a 32D model and compared the binding of C/EBPα with a C-terminal DNA-binding mutant form that was previously found in AML[35]. Moreover, we demonstrate that combining ChIP-on-chip with gene expression data reveals the effects of direct target interactions. More specifically, we identified a subset of putative C/EBPα target genes that were differentially expressed in *CEBPA*^{silenced} AMLs compared to *CEBPA* expressing AML samples. Importantly, the overlapping list of gene promoters was highly enriched for C/EBPα binding sites. In other words, this list may contain genes that are bound by wild-type and not by mutant C/EBPα through interaction with "classical" C/EBPα binding sites. It is predicted that differential gene expression profiles between *CEBPA* silenced and other AML samples is at least partially causing the effects. The overlap was highly significant, meaning that determining actual binding of C/EBPα to promoters in a murine cell line model, provides insight into the effects of C/EBPα in human AML cells. It is important to realize that the use of array-data, such as ChIP-on-chip is limited to the promoters defined on the chip. A large number of regulatory regions lie outside the promoter regions and sometimes at long distances from the genes targeted by these factors. It will therefore be of interest to carry out deep sequencing rather than ChIP hybridizations following chromatin-IP of wild-type versus mutant C/EBPα binding in 32D model. Moreover, we compared binding of C/EBPα in mouse models, and it is known that there are differences between mouse and human myeloid cells. Our data show that it is possible to identify direct targets of transcription factors in leukemia samples.

We overlaid the detected direct binding targets of C/EBPα in 32D cells, with the differentially expressed genes in the *CEBPA*^{silenced} group and identified 49 overlapping genes as putative direct targets. Among these 49 genes, 25 were downregulated and 24 were upregulated in the primary *CEBPA*^{silenced} AML group. The 25 downregulated genes represent putative targets of C/EBPα that are activated under normal conditions. The 24 upregulated genes are highly enriched for pathways in T-cell development and are repressed in wild-type *CEBPA* expressing cells and activated when *CEBPA* was silenced. This suggests that the transcription factor C/EBPα may also acts as a repressor of gene transcription. Transcription factor binding site analysis showed that 20 of the 24 upregulated genes contain the CEBP consensus sequence. Moreover, other known transcription factor binding sites, e.g. PU.1[185,186], SP1[187] and E2F[188,189] were found as well in the promoter regions (all located within 2kb). Transcription factors recognizing these binding sites were previously identified to interact with C/EBPα. Interestingly, E2F has been reported to be involved in the downregulation of genes[188,189]. In fact, it was recently shown that *CEBPA* represses *Cebpg* expression by affecting E2F1 transcriptional activity[184]. In luciferase reporter assays using the C/EBPγ proximal promoter it was demonstrated that both C/EBPα isoforms, i.e. the p30 and the p42 isoform can repress C/EBPγ transactivation. Our studies are also in line with our previous report, showing that multiple T-cell related genes, e.g. *CD7* or *LCK* are downregulated by C/EBPα[195]. Similar to what we have demonstrated here, these genes were strongly upregulated in mouse bone marrow LSK cells in which *Cebpa* was knocked out. Upon reintroduction of *Cebpa,* the expression of these genes rapidly declined[195]. Taken together, these findings suggest that C/EBPα can not only act as an activator, but also as a repressor of transcription. Whether C/EBPα functions as an activator or a repressor is most likely determined by distinct complexes containing additional transcriptional repressors and activators. Both functions seem essential for a proper balance between proliferation and differentiation of primitive progenitor cells in the bone marrow.

AUTHORSHIP

Contribution: E.T. performed research, data analyses, data interpretation, statistical analyses and drafted the manuscript. R.A. performed biological experiments and manuscript editing. J.d.R., and M.J.T.R. participated in the statistical discussion, data interpretation and manuscript editing. B.J.W. performed biological experiments and manuscript writing. C.A.E.J performed biological experiments. R.D. contributed to the biological design of the study, data interpretation and manuscript editing. E.T., R.D., B.J.W., J.d.R., M.J.T.R. and R.A. participated in the discussion of the results. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This research was performed within the framework of CTMM, the Center for Translational Molecular Medicine, project BioCHIP (grant 03O-102).

FIGURE LEGENDS

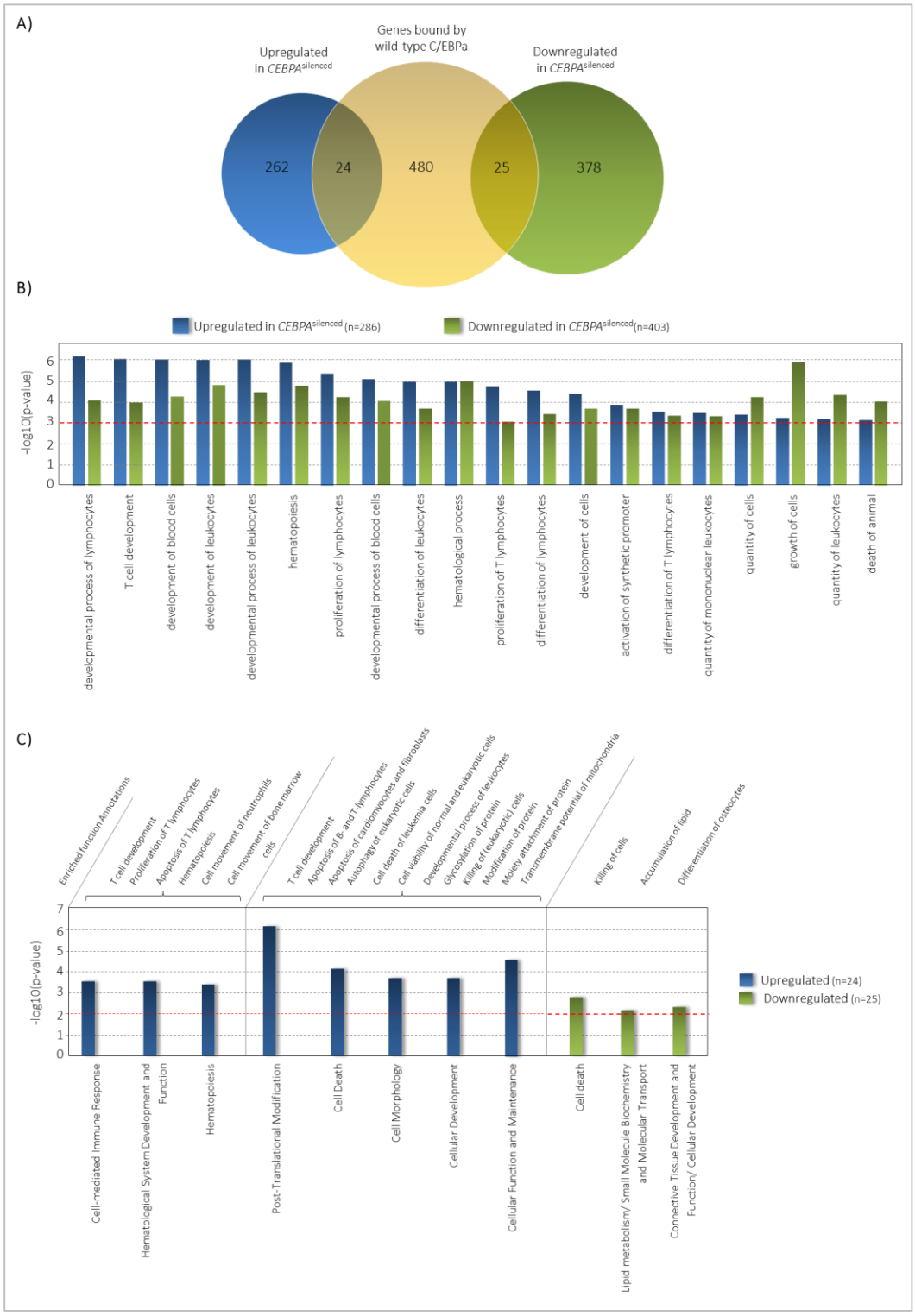


Figure 1. Overview *CEBPA*^{silenced} targets genes and enriched pathways. (A) Genes that are bound by wild-type C/EBPα in the 32D model-system (n=529, yellow circle), and that overlap with the *CEBPA*^{silenced} leukemias when compared to

normal CD34^{+} and versus the other AMLs. The blue circle indicates the upregulated (n=286) genes from which 24 genes are bound by C/EBPα in the proximal promoter region. The green circle indicates the downregulated (n=403) genes from which 25 genes are bound by C/EBPα in the proximal promoter region. (B) Graphical representation of enriched pathways (shown in bar graphs) that overlapped between the 286 upregulated and 403 downregulated genes. (C) Graphical representation of the enriched categories for the 24 upregulated and 25 downregulated genes that were bound by C/EBPα in the 32D model-system.

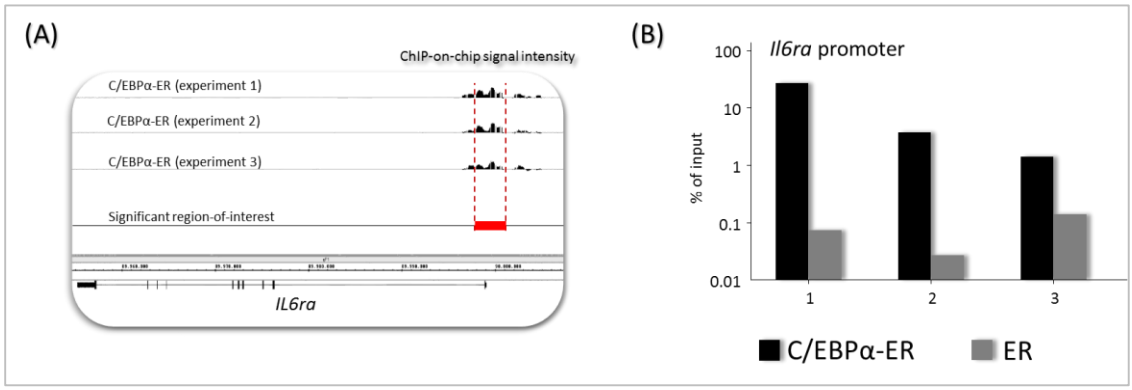


Figure 2. C/EBPα binding for Il6ra. (A) Affymetrix genome browser illustrates relative probe intensity signal with vertical bars of three C/EBPα-ER clones compared to C/EBPα-MUTANT-ER. The red rectangle illustrate the significantly detected binding region in the promoter region of Il6ra in all C/EBPα-ER experiments. (B) Bar plots depicts the binding of C/EBPα-ER and ER in three experiments using a quantitative ChIP-Q-PCR.

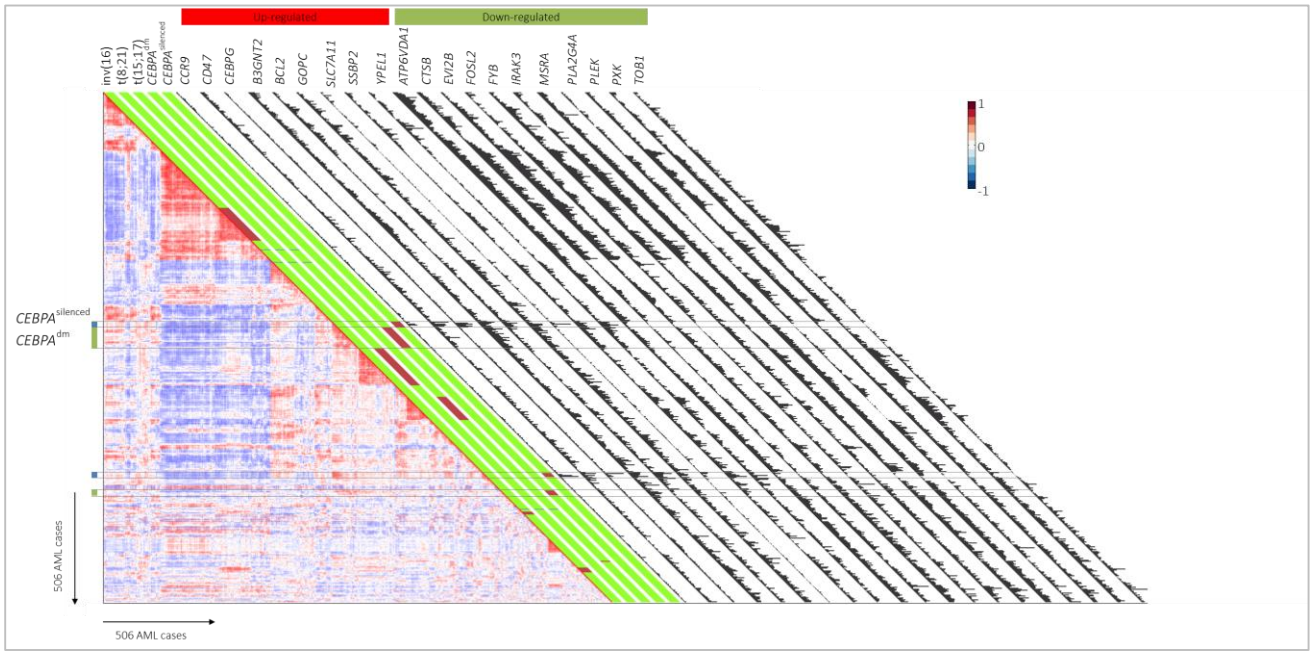


Figure 3. Aberrant mRNA expression levels of C/EBPα target genes in the *CEBPA*^{silenced} group. Pairwise correlations between the 506 AML cases (A). The cells in the visualization are colored by Pearson correlations values, depicting

higher positive (red) or negative (blue) correlations, as indicated by the scale bar. The cytogenetical groups; inv(16), t(8;21), t(15;17), together with *CEBPA*^{dm} and *CEBPA*^{silenced} cases are depicted on the diagonal with a red colored bar. For 9 upregulated and 11 downregulated genes we illustrate the expression profile on the diagonal as black bars that were detected as significantly differential expressed in human AML and bound by C/EBPα in the 32D model-system.

Table 1. Candidate target genes for *CEBPA*^{silenced}, prior the scenarios of the 32D model system

Group	Overlap with wild-type C/EBPα in 32D cells (n=529)	
	Upregulated	**Downregulated**
***CEBPA*^{silenced}**	*ACAD8, ATP6V1B2, B3GNT2, B4GALT1, BCL2*, CASP1, CCR9*, CD47, CEBPG*, CEPT1, FBXW2, GOPC, HMGCR, MAP2K4, MRPL41, NIN, NMT1, OGT, SEPX1, SLC7A11, SSBP2, STK38, UBTD1, YPEL1*	*ACSL1, ARSB, ATP6V0A1, COL4A3BP, CTSB, DEGS1, EVI2B, FOSL2, FYB, GOSR1, H2AFJ, HIPK1, HSP90B1, IRAK3, LIN7C, MBTD1, MSRA, MYCT1, PLA2G4A, PLEK, PPM1A, PXK, RANBP9, STX11, TOB1*
(689 genes)	(49 genes, $P = 1x10^{-7}$)	

Table 1. Candidate target genes for *CEBPA*^{silenced}, prior the scenarios of the 32D model-system. Human gene-symbols are listed that overlapped with the candidate C/EBPα target genes from the 32D model and were called differential expressed for C/EBPαsilenced group. The *P-value* indicates the probability of detecting this overlap by chance. Bolded gene-symbols with an asterisk are previously reported in literature to interact with C/EBPα.

SUPPORTING MATERIAL

C/EBPα-ER cells differentiate upon E_2 treatment while C/EBPα-mutant-ER cells show impaired differentiation

Stable 32D clones expressing a fusion protein of murine C/EBPα and the ligand-binding domain (LBD) of the human estrogen receptor alpha (ERá) (C/EBPα-ER) were generated. 32D cells proliferate when stimulated with IL-3 and differentiate into granulocytes when induced by G-CSF[183]. Upon addition of E_2, C/EBPα-ER cells morphologically differentiated into mature granulocytes even in the presence of IL-3 (Figure S3A), while their proliferation rate dropped (Figure S3B). This effect appeared even more pronounced when the cells were treated with G-CSF plus E2. (Figure S3, panel C-D). As expected, control cells only expressing the LBD of ERá (ER) did not show a response to E_2 (Figure S3B and data not shown).

32D cells expressing a mutant C/EBPα-ER fusion with an insertion of 6 amino acids in the bZIP domain (C/EBPα-mutant-ER) as previously found in a human AML patient[36] showed less pronounced inhibition of proliferation upon treatment with E_2 in the presence of IL3. Morphological differentiation of these cells could hardly be detected under IL3 plus E_2 conditions (Figure S3A-B). In the presence of G-CSF, C/EBPα-mutant-ER cells demonstrated delayed differentiation with the suggestion of a partial block (Figure S3C).

Candidate C/EBPα targets by using ChIP-on-chip promoter arrays

For the detection of candidate target genes that are bound by wild-type protein C/EBPα, we used HAT[182] and compared C/EBPα-ER clones (n=3) to C/EBPα-mutant-ER clones (n=2). The C/EBPα-mutant-ER clones cannot bind to the DNA compared to the ER clones, and eliminates DNA-binding effects caused by ER. In any of the wild-type C/EBPα-ER clones versus C/EBPα-mutant-ER, we detected 2732 statistically significant binding regions sized between 104 and 9928 nucleotides (median 823nt); 80% (2185) of these binding regions were detected in two or more clones whereas the 2732 binding regions could be mapped to 964 unique genes within 2Kb of the transcriptional-start-site.

To investigate the enrichment of transcription factor binding motifs for the detected binding regions, we utilized F-MATCH[163,164]. This resulted in the detection of six highly enriched TFBSs (P<1x10^{-28}) from which five are known to be C/EBPα consensus sites. Furthermore, TFBSs were four times more observed than in the background set (fold-increase > 4) (Table S5). This indicates high specificity of C/EBPα binding in the detected regions.

Among the mapped genes we observed genes that have previously been reported to be regulated by C/EBPα, such as *Il6ra*[82], *C3*[52], *Hp*[52], *Mpo*[196], *Myc*[197] and *Sfpi1*[198]. The gene encoding the IL-6 receptor alpha has previously been identified as a critical downstream target of C/EBPα. These latter findings give confidence, that among the putative novel C/EBPα target genes, critical players may be hidden, important in neutrophilic development and consequently in myeloid transformation.

Plasmids

A pBabe-Cebpa-ER fusion construct and pBabe-ER construct were provided by Dr. Daniel Tenen (Harvard Institutes of Medicine, Boston, MA, USA). The Cebpa-ER fusion gene was re-cloned into the pLNCX-neo vector using polymerase chain reaction (PCR) amplification with primers that included appropriate restriction sites, i.e. *Hpa*I and *Cla*I. Primer sequences were: fw 5'-GTA CGT TAA CAG GAA TTC GCG CCA CCA TGG A-3' and rev 5'-AGG AAT CGA TCT CTC AGA CTG TGG CAG GGA A-3'. A mutant construct was generated using the QuikChange site-directed mutagenesis kit (Stratagene, La Jolla, CA, USA), using the following oligos (insertion mutation is underlined): 5'-GAT AAA GCC AAA CAA CGT AAT GTG GAC AAG CAG CGC AAC GTG GAG-3' (sense) and 5'-CTC CAC GTT GCG CTG CTT GTC CAC ATT ACG TTG TTT GGC TTT ATC-3' (anti-sense). A pLNCX-ER construct was generated with primers fw 5'-TAT GGT TAA CAT CTG CTG GAG ACA TGA GAG CT-3' and the reverse primer used for the Cebpa-ER construct. The cloned constructs and site of inserted mutation were nucleotide-sequenced.

Figure Legends

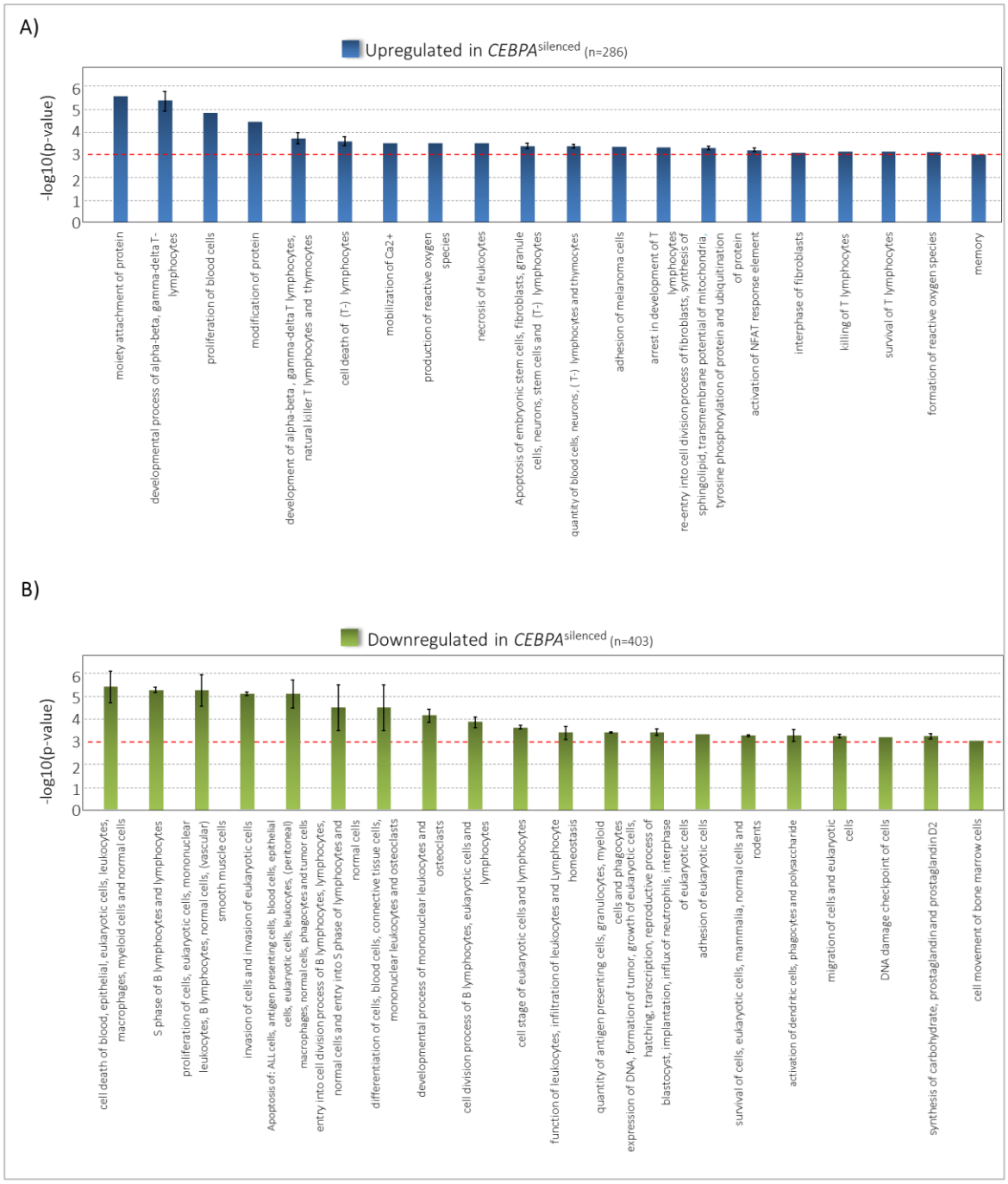


Figure S1. Enriched functional pathways for the 286 upregulated and 403 downregulated genes that are detected in the *CEBPA*^{silenced} cases. Graphical representation of the functional pathways (shown in bar graphs) that are significantly enriched in (A) 286 upregulated genes, and (B) 403 downregulated genes for the *CEBPA*^{silenced} AMLs.

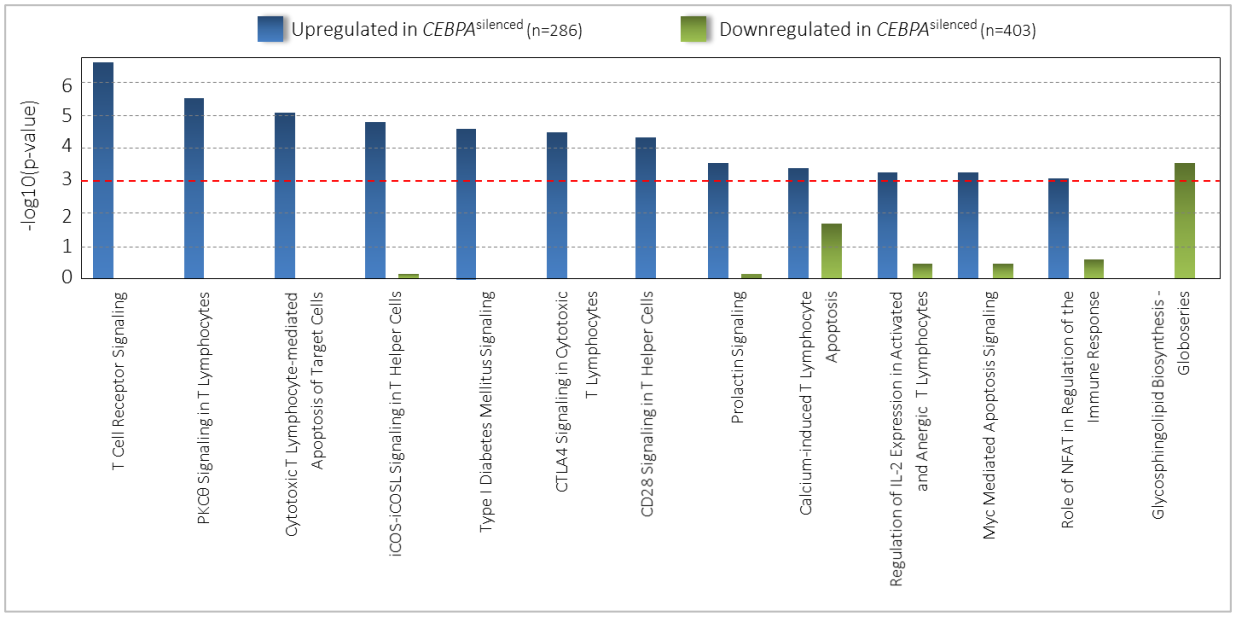


Figure S2. Enriched canonical pathways for the 286 upregulated and 403 downregulated genes that are detected in the *CEBPA*^{silenced} cases. Graphical representation of the canonical pathways (shown in bar graphs) that are significantly enriched in (A) 286 upregulated genes, and (B) 403 downregulated genes for the *CEBPA*^{silenced} AMLs.

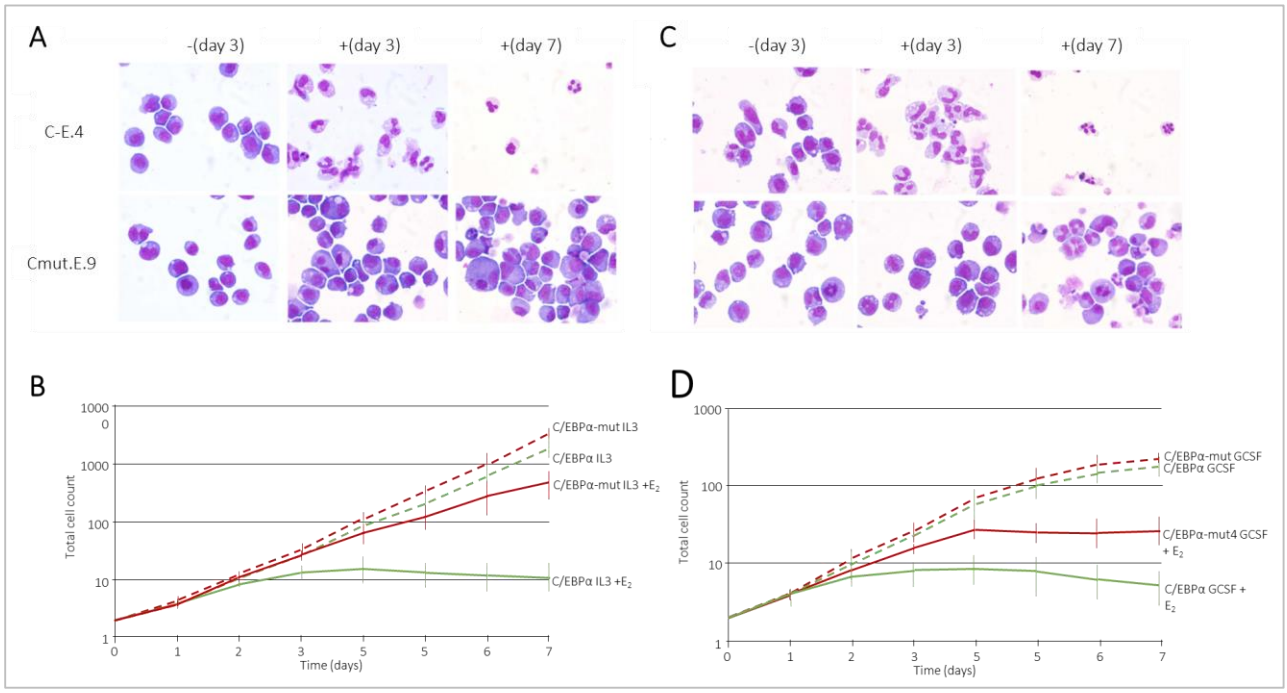


Figure S3. Differentiation of C/EBPα among IL3 conditions. (A) Representative cytospins of C/EBPα-ER cells (clone C-E.4) and C/EBPα-mutant-ER cells (clone Cmut-E.9) in the absence (-) and presence (+) of E_2 for 3 or 7 days. The cells were cultured in the presence of IL3. (B) Proliferation curves of C/EBPα-ER cells (C/EBPα) and C/EBPα-mutant-ER cells

(C/EBPα-mutant) stimulated with IL3 alone (IL3 -) or IL3 in combination with E_2 (IL3 +est). The graph represents the average and standard deviations for the individual clones for the cell types. (C) Representative cytospins of C/EBPα-ER cells (clone C-E.4) and C/EBPα-mutant-ER cells (clone Cmut-E.9) in the absence (-) and presence (+) of E_2 for 3 or 7 days. The cells were cultured in the presence of G-CSF. (D) Proliferation curves of C/EBPα-ER cells (C/EBPα) and C/EBPα-mutant-ER cells (EBPα-mutant) stimulated with G-CSF alone (G-CSF -) or IL3 in combination with E_2 (G-CSF +est). The graph represents the average and standard deviations for the individual clones for each of the cell types.

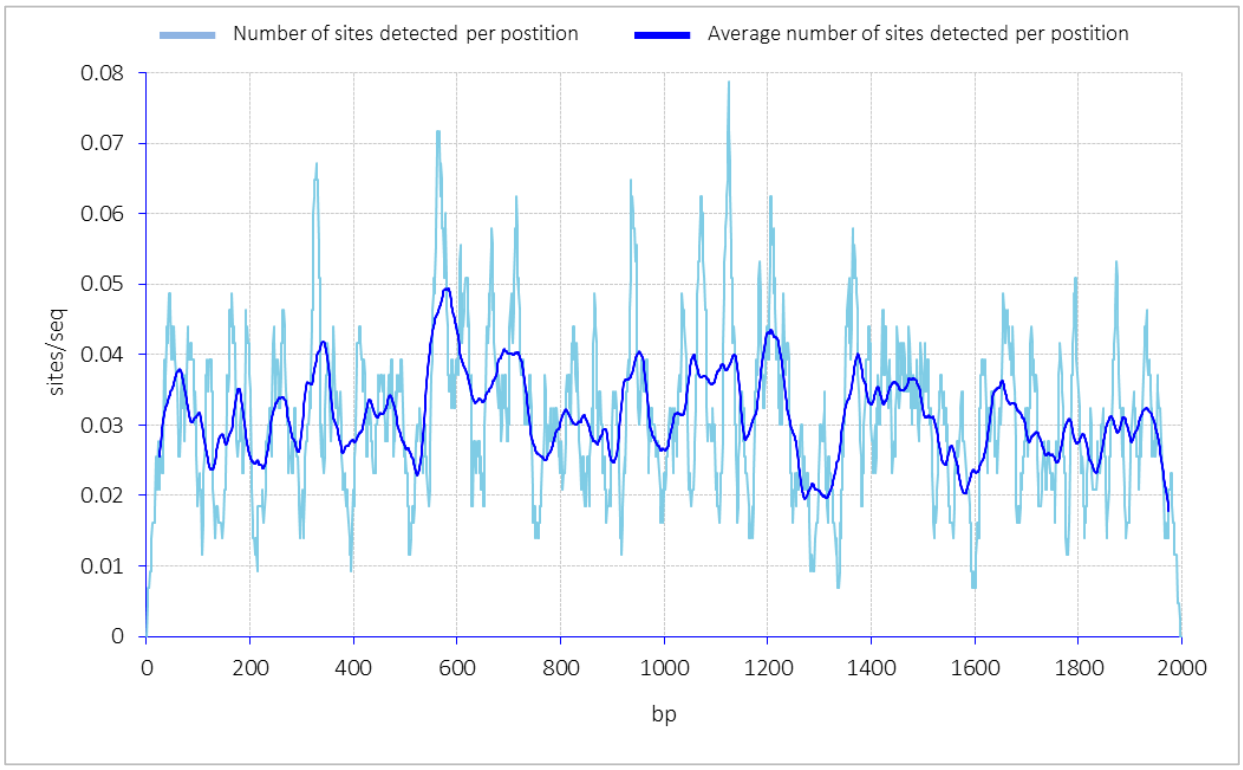


Figure S4. Distribution of the detected CEBP consensus sites. Graphical representation of the genomic distance (base pairs; bp) for the CEBP consensus sites that are detected in the proximal promoter regions of the 529 genes. Dark blue is the average number of sites detected per position whereas light blue the actual number of sites per position.

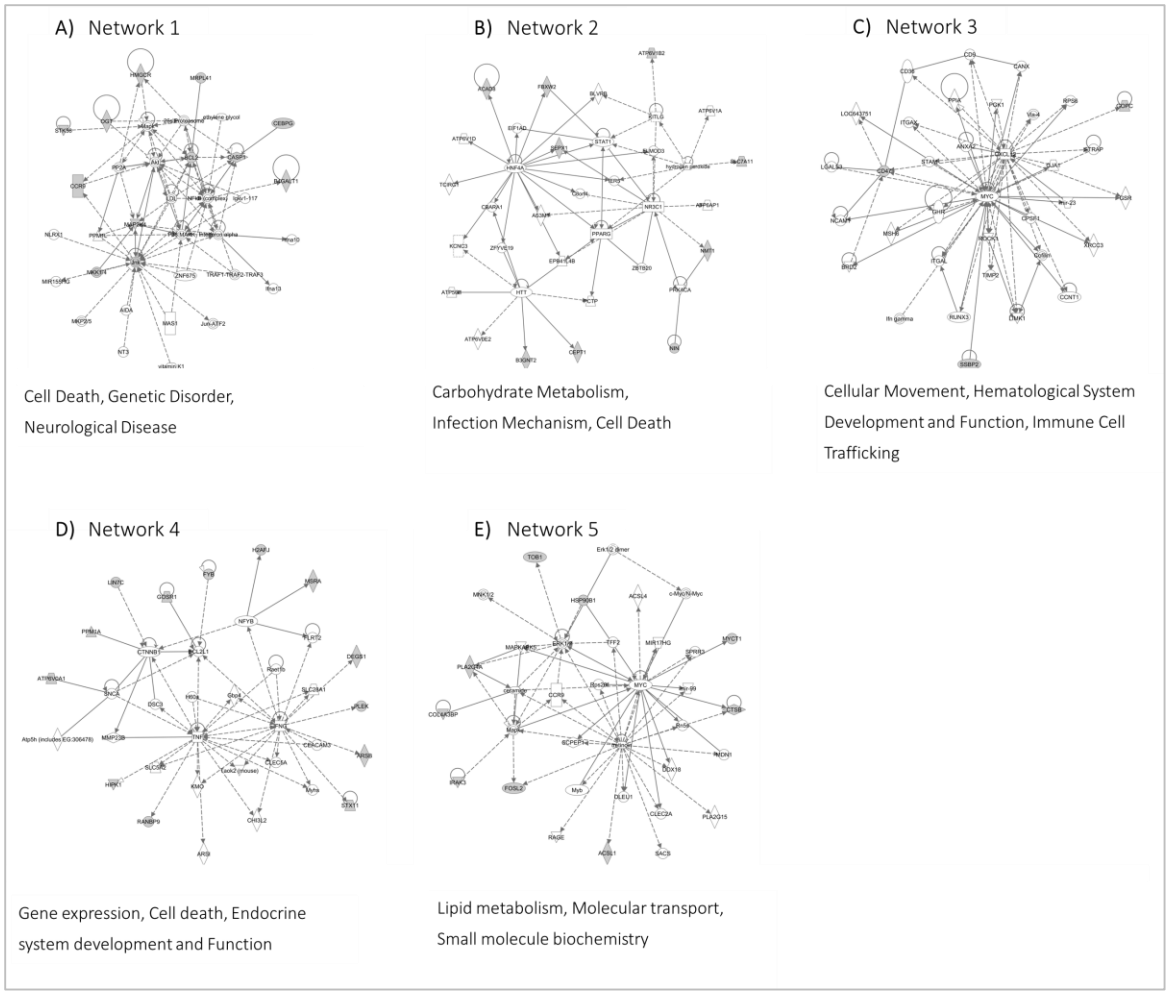


Figure S5. Enriched networks for the 24 upregulated and 25 downregulated genes that are detected in the *CEBPA*^{silenced} cases. Illustration of networks that are enriched using IPA for the 24 upregulated genes (panel A, B and C), and 25 downregulated genes (Panel D and E). The genes are dark grey coloured in the networks and are bound by wild-type C/EBPα in the 32D model system and differentially expressed in the *CEBPA*^{silenced} group.

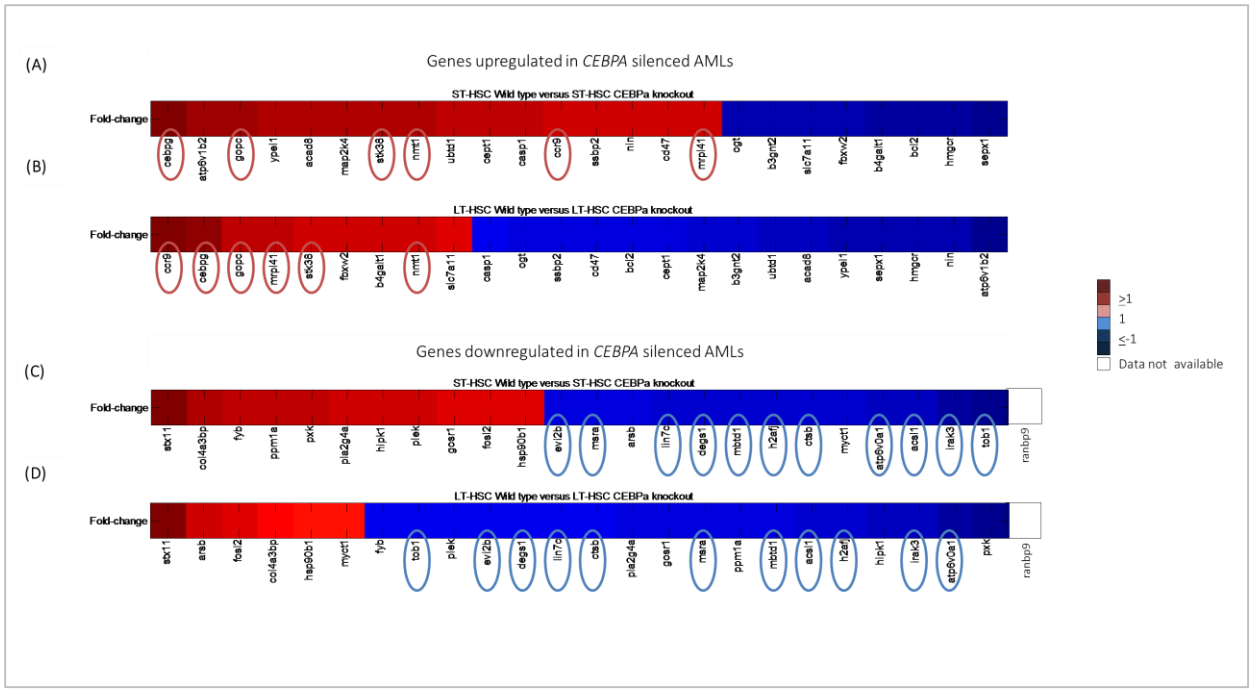


Figure S6. Validation with LT-HST and ST-HST *Cebpa*-KO mouse and wild-type 32D C/EBPα-ER cell line. The 24 upregulated genes among the *CEBPA*^{silenced} leukemias are overlaid with (A) the genes that showed a fold-change > 1 in the ST-HSC wild-type versus ST-HSC knockout expression levels and (B) the genes that showed a fold-change > 1 in the LT-HSC wild-type versus LT-HSC knockout expression levels. The red-circled genes illustrates the overlap between A and B. The 25 downregulated genes among the *CEBPA* silenced leukemias are overlaid with (C) the genes that showed a fold-change < 1 in the ST-HSC wild-type versus ST-HSC knockout expression levels and (B) the genes that showed a fold-change < 1 in the LT-HSC wild-type versus LT-HSC knockout expression levels. The blue-circled genes illustrates the overlap between C and D.

Pathways detected in:	Function Annotation	Regulation	*P*-value	Genes	Nr. of genes
Both upregulated and downregulated genes	T cell development	DOWN	9.72E-05	*BCL2L11,BCL3,CD40,CD74,CDK2,CEBPA,CHEK1,CST3,E2F1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA,IL18,JUNB,MBD2,MCL1,N*	33
		UP	7.45E-07	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,FAS,GRAP2,ITGA4,LCK,MAP2K4,NC*	30
	activation of synthetic promoter	DOWN	2.05E-04	*BEX2,CD40,CD74,CITED4,CYTIP,DACH1,E2F1,FOXO1,GFI1,HBEGF,HSP90B1,IL18,JUNB,LTBR,NFATC2,NFE2L2,NR4A2,PLA2G4A,PPP3C*	32
		UP	1.28E-04	*ANGPT2,APC,ATM,BCL2,CBL,DLX1,DUSP1,FAS,JAK2,JARID2,LCK,MACF1,MAP2K4,MLX,MTF1,MYCN,NOTCH1,PRKCQ,SART3,SH2D1A*	25
	death of animal	DOWN	8.24E-05	*ADRB2,APLP2,ATP2B1,BCL2L11,BLMH,CD40,CDH2,CDK2,CEBPD,CHEK1,CTSB,DACH1,FOXO1,GFI1,GNA11,GNAS,HBEGF,HIPK1,HMGI*	49
		UP	7.17E-04	*ABL2,ADCYAP1,ANGPT2,APAF1,APC,ATF2,ATM,BCL2,BUB3,CASP1,CBL,CEBPG,CUL7,CYBB,DLX3,DUSP1,EPHB3,FAS,GPX1,HMGCR,I*	34
	development of blood cells	DOWN	4.92E-05	*BCL2L11,BCL3,BST1,CAST,CD40,CD74,CDK2,CEBPA,CHEK1,CSF1R,CST3,E2F1,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA*	39
		UP	7.76E-07	*ADCYAP1,ANGPT2,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,FAS,GRAP2,ITGA4,JAK2,LC*	34
	development of cells	DOWN	1.92E-04	*ANXA5,BASP1,BCL2L11,BCL3,BST1,CAST,CD40,CD74,CDH2,CDK2,CEBPA,CHEK1,COL4A3BP,CSF1R,CSNK2A2,CST3,CTNNA1,CTSB,E2F*	63
		UP	3.71E-05	*ABL2,ADCYAP1,ANGPT2,APAF1,APC,ATM,ATP7A,B4GALT1,BCL11B,BCL2,CASP1,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,CLDN5,DEF*	49
	development of leukocytes	DOWN	1.39E-05	*BCL2L11,BCL3,BST1,CAST,CD40,CD74,CDK2,CEBPA,CHEK1,CSF1R,CST3,E2F1,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA*	38
		UP	8.20E-07	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,FAS,GRAP2,ITGA4,JAK2,LCK,MAP2*	32
	developmental process of blood cells	DOWN	8.24E-05	*BCL2L11,BCL3,BST1,CAST,CD40,CD74,CDK2,CEBPA,CEBPD,CHEK1,CSF1R,CST3,E2F1,ELANE,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,I*	48
		UP	7.59E-06	*ADCYAP1,ANGPT2,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CBX2,CCR9,CD247,CD3D,CD3G,CD47,CD7,CLCF1,EPHB6,FAS,GI*	39
	developmental process of leukocytes	DOWN	2.94E-05	*BCL2L11,BCL3,BST1,CAST,CD40,CD74,CDK2,CEBPA,CEBPD,CHEK1,CSF1R,CST3,E2F1,ELANE,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,I*	45
		UP	8.45E-07	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CBX2,CCR9,CD247,CD3D,CD3G,CD47,CD7,CLCF1,EPHB6,FAS,GRAP2,ITG*	38
	developmental process of lymphocytes	DOWN	7.85E-05	*BCL2L11,BCL3,BST1,CD40,CD74,CDK2,CEBPA,CHEK1,CST3,E2F1,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,IL13RA1,IL15RA,IL1*	38
		UP	5.55E-07	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CBX2,CCR9,CD247,CD3D,CD3G,CD47,CD7,CLCF1,EPHB6,FAS,GRAP2,ITG*	34
	differentiation of T lymphocytes	DOWN	4.26E-04	*BCL3,CD40,CD74,CEBPA,FZD8,GFI1,HSP90B1,ID1,IFNGR1,IL18,JUNB,MBD2,MYLK,NFATC2,PRKCI,PTGS2,RAG1,RIPK2,TLR2,TNFSF13E*	20
		UP	2.85E-04	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,ITGA4,LCK,NOTCH1,PRKCQ,SH2D1A,STAT5B,TCF7,TOX,ZAP70*	16
	differentiation of leukocytes	DOWN	1.97E-04	*BCL3,CD40,CD74,CEBPA,CEBPD,CSF1R,FZD8,GFI1,HSP90B1,ID1,IFNGR1,IL13RA1,IL15RA,IL18,JAG1,JUNB,KLF4,LTBR,MBD2,MYLK,ND*	30
		UP	9.80E-06	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CD3D,CD47,CLCF1,FAS,ITGA4,JAK2,LCK,MYCN,NFYA,NOTCH1,OGT,PRKCQ,SH2D1A*	26
	differentiation of lymphocytes	DOWN	3.75E-04	*BCL3,CD40,CD74,CEBPA,FZD8,GFI1,HSP90B1,ID1,IFNGR1,IL13RA1,IL15RA,IL18,JUNB,MBD2,MYLK,NDFIP1,NFATC2,PRKCI,PTGS2,RA*	24
		UP	2.74E-05	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CD3D,CLCF1,FAS,ITGA4,LCK,MYCN,NOTCH1,PRKCQ,SH2D1A,STAT5B,TCF7,TOX,TR*	21
	growth of cells	DOWN	1.15E-06	*ACTG1,ACTN1,ADAM15,BCL2L11,BRF1,CAST,CD40,CDK2,CEBPA,CEBPD,CGREF1,CHEK1,CHFR,CREG1,CSF1R,CTSB,DACH1,DEGS1,E2F*	83
		UP	5.47E-04	*ABL2,AIMP1,APBA3,APC,ARL6IP5,ATF2,ATM,BCL2,CASP1,CBL,CBX2,CIRBP,COL2A1,CUL7,CYBA,DGKZ,DUSP1,EPHB3,EPHB6,ERCC3,I*	53
	hematological process	DOWN	9.49E-06	*ANXA5,BCL2L11,BCL3,BST1,CAST,CD40,CD74,CDK2,CEBPA,CEBPD,CHEK1,CSF1R,CST3,E2F1,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,*	50
		UP	9.73E-06	*ADCYAP1,ANGPT2,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,CLCF1,DUSP1,FAS,GRAP2,*	38
	hematopoiesis	DOWN	1.55E-05	*BCL2L11,BCL3,BST1,CAST,CD40,CD74,CDK2,CEBPA,CEBPD,CHEK1,CSF1R,CST3,E2F1,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,*	45
		UP	1.23E-06	*ADCYAP1,ANGPT2,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,CLCF1,FAS,GRAP2,ITGA4,J*	37
	proliferation of T lymphocytes	DOWN	8.57E-04	*ABCG1,ADRB2,CD40,CD74,CDK2,CTSZ,FYB,GFI1,HMGB1,HSP90B1,IFNGR1,IL18,IRS2,JAG1,MVP,NDFIP1,NFATC2,PTGS2,RAG1,RIPK2,*	25
		UP	1.68E-05	*ADCYAP1,ATM,B3GNT2,BCL2,CBL,CD247,CD3G,CD47,DUSP1,EPHB6,FAS,GCNT1,GRAP2,LCK,MAP2K4,NOTCH1,PRKCQ,SH2D1A,SPH*	23
	proliferation of lymphocytes	DOWN	5.68E-05	*ABCG1,ADRB2,BCL2L11,CD40,CD74,CDK2,CTSZ,FYB,GFI1,HMGB1,HSP90B1,IFNGR1,IL13RA1,IL15RA,IL18,IRS2,JAG1,JUNB,KLF4,KLF9,*	33
		UP	3.96E-06	*ADCYAP1,ATM,B3GNT2,BCL2,CBL,CD247,CD3G,CD47,CLCF1,DUSP1,EPHB6,FAS,GCNT1,GRAP,GRAP2,LCK,MAP2K4,MYCN,NOTCH1,F*	28
	quantity of cells	DOWN	5.55E-05	*APLP2,ARL4A,BCL2L11,BHLHE40,BST1,CD40,CDK2,CEBPA,CSF1R,CTSB,CYTIP,DNMT1,E2F1,FOSL2,FOXO1,GFI1,GNG7,HBEGF,HEXB,H*	54
		UP	3.66E-04	*ADCYAP1,ANGPT2,APAF1,ATF2,ATM,ATRX,BCL2,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,CYBB,DGKZ,DLX1,DUSP1,FAS,GCNT1,GPX*	38
	quantity of leukocytes	DOWN	4.36E-05	*BCL2L11,BST1,CD40,CEBPA,CSF1R,CYTIP,DNMT1,E2F1,GFI1,HBEGF,HEXB,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA,IL18,KLF4,LTBR,MXD*	32
		UP	6.03E-04	*ATM,BCL2,CCR9,CD247,CD3D,CD3G,CD47,CD7,FAS,GCNT1,GRAP2,JAK2,JARID2,LCK,NOTCH1,PRF1,SIGIRR,STAT5B,TCF7,TOX,ZAP70,*	22
	quantity of mononuclear leukocytes	DOWN	4.75E-04	*BCL2L11,BST1,CD40,CYTIP,DNMT1,E2F1,GFI1,HEXB,HSP90B1,ID1,IFNGR1,IL15RA,IL18,KLF4,LTBR,NFATC2,PRKCD,PROK2,PRTN3,RAG*	24
		UP	3.01E-04	*ATM,BCL2,CCR9,CD247,CD3D,CD3G,CD7,FAS,GCNT1,GRAP2,JAK2,JARID2,LCK,NOTCH1,PRF1,STAT5B,TCF7,TOX,ZAP70*	19

Upregulated genes	activation of NFAT response element	UP	5.90E-04	*CBL,GRAP2,NOTCH1,ZAP70*	4
	activation of T lymphocytes	UP	6.48E-04	*BCL2,CD247,CD3D,CD3G,CD47,CD7,DUSP1,FAS,LCK,PRF1,PRKCQ,SLA2,SPHK2,STAT5B*	14
	adhesion of melanoma cells	UP	4.53E-04	*CD47,ITGA4,MCAM*	3
	apoptosis of embryonic stem cells	UP	5.90E-04	*APAF1,BCL2,FAS,MAP2K4*	4
	apoptosis of fibroblasts	UP	5.00E-04	*ABL2,APAF1,ATF2,BCL2,CASP1,DUSP1,FAS,MAP2K4,NOTCH1,SKP2,SPHK2*	11
	apoptosis of granule cells	UP	9.14E-04	*ADCYAP1,ATF2,BCL2,FAS,NFATC4*	5
	apoptosis of lymphocytes	UP	6.71E-04	*ADCYAP1,AIMP1,ATM,BCL2,CASP1,CD247,CD47,FAS,GRAP2,LCK,MAP2K4,PRF1,STAT5B,TCF7*	14
	apoptosis of neurons	UP	1.80E-04	*ADCYAP1,APAF1,ATF2,ATM,ATRX,BCL2,CASP1,CYBA,DLX1,DUSP1,FAS,GPX1,MAP2K4,MECP2,MYCN,NFATC4*	16
	apoptosis of stem cells	UP	3.36E-04	*APAF1,ATM,BCL2,FAS,MAP2K4*	5
	apoptosis of T lymphocytes	UP	2.63E-04	*ADCYAP1,ATM,BCL2,CASP1,CD247,CD47,FAS,GRAP2,LCK,MAP2K4,PRF1,STAT5B,TCF7*	13
	arrest in development of T lymphocytes	UP	4.53E-04	*CBL,CD3D,LCK*	3
	cell death of lymphocytes	UP	3.38E-04	*ADCYAP1,AIMP1,APAF1,ATM,BCL2,CASP1,CD247,CD47,CD7,FAS,GRAP2,LCK,MAP2K4,PRF1,STAT5B,TCF7*	16
	cell death of T lymphocytes	UP	9.70E-05	*ADCYAP1,APAF1,ATM,BCL2,CASP1,CD247,CD47,CD7,FAS,GRAP2,LCK,MAP2K4,PRF1,STAT5B,TCF7*	15
	development of alpha-beta T lymphocytes	UP	1.12E-04	*CD3G,LCK,NOTCH1*	3
	development of gamma-delta T lymphocytes	UP	5.69E-05	*CD3D,CD3G,NOTCH1*	3
	development of natural killer T lymphocytes	UP	8.66E-04	*NOTCH1,PRKCQ,SH2D1A*	3
	development of thymocytes	UP	1.65E-04	*BCL2,CCR9,CD3D,CD3G,TCF7,TRAT1,ZAP70*	7
	developmental process of alpha-beta T lymphocytes	UP	2.56E-06	*BCL11B,CD3G,LCK,NOTCH1,TCF7,ZAP70*	6
	developmental process of gamma-delta T lymphocytes	UP	9.62E-06	*CD3D,CD3G,LCK,NOTCH1,STAT5B*	5
	developmental process of T lymphocytes	UP	1.64E-07	*ADCYAP1,APAF1,APC,ATM,ATP7A,BCL11B,BCL2,CASP1,CBL,CBX2,CCR9,CD247,CD3D,CD3G,CD47,CD7,EPHB6,FAS,GRAP2,ITGA4,LCK*	32
	formation of reactive oxygen species	UP	7.28E-04	*CYBA,CYP2E1,FAS,JAK2,NCF4*	5
	interphase of fibroblasts	UP	7.95E-04	*ATF2,ATM,CBX2,CD247,MYCN,SKP2*	6
	killing of T lymphocytes	UP	7.05E-04	*CD47,FAS,PRF1,SH2D1A*	4
	memory	UP	9.51E-04	*ABL2,ADCYAP1,BCL2,CASP1,CD47,CYBB,MECP2,NOTCH1,SIGIRR*	9
	mobilization of Ca2+	UP	2.87E-04	*ADCYAP1,CBL,CD247,CD3D,CD3G,LCK,PRKCQ,SPHK2,TRAT1,ZAP70*	10
	modification of protein	UP	3.33E-05	*ABL2,ALG9,ATM,B3GNT2,B4GALT1,BCL2,CAND1,CASP1,CBL,CD247,CD47,CD7,CYBB,DUSP1,EPHB6,ERCC3,FAS,FBXW2,GCNT1,GPX1,*	42
	moiety attachment of protein	UP	2.47E-06	*ABL2,ALG9,ATM,B3GNT2,B4GALT1,BCL2,CAND1,CBL,CD247,CD47,CD7,DUSP1,EPHB6,ERCC3,FAS,FBXW2,GCNT1,GPX1,JAK2,KLHL9,*	38
	necrosis of leukocytes	UP	3.06E-04	*BCL2,CASP1,FAS*	3
	production of reactive oxygen species	UP	3.26E-04	*ADCYAP1,ANGPT2,BCL2,CD47,CYBB,CYP2E1,FAS,FDXR,MAP2K4,PRF1,PRKCQ,ZAP70*	12
	proliferation of blood cells	UP	1.34E-05	*ADCYAP1,ATM,B3GNT2,BCL2,CBL,CD247,CD3G,CD47,CLCF1,DUSP1,EPHB6,FAS,GCNT1,GRAP,GRAP2,JAK2,LCK,MAP2K4,MYCN,NOT*	29
	quantity of blood cells	UP	8.79E-04	*ATM,BCL2,CBL,CCR9,CD247,CD3D,CD3G,CD47,CD7,FAS,GCNT1,GRAP2,JAK2,JARID2,LCK,NOTCH1,PRF1,SIGIRR,STAT5B,TCF7,TOX,ZA*	23
	quantity of lymphocytes	UP	4.82E-04	*ATM,BCL2,CCR9,CD247,CD3D,CD3G,CD7,FAS,GCNT1,GRAP2,JAK2,LCK,NOTCH1,PRF1,STAT5B,TCF7,TOX,ZAP70*	18
	quantity of neurons	UP	4.15E-04	*ADCYAP1,ANGPT2,APAF1,ATM,BCL2,CYBB,DLX1,DUSP1,GPX1,MECP2,NOTCH1*	11
	quantity of T lymphocytes	UP	2.16E-04	*ATM,BCL2,CCR9,CD247,CD3D,CD3G,CD7,FAS,GRAP2,JAK2,LCK,NOTCH1,STAT5B,TOX,ZAP70*	15
	quantity of thymocytes	UP	3.95E-04	*ATM,CCR9,CD247,FAS,GRAP2,LCK,NOTCH1,TOX*	8
	re-entry into cell division process of fibroblasts	UP	8.66E-04	*ATM,BCL2,MYCN*	3
	survival of T lymphocytes	UP	9.26E-04	*ADCYAP1,BCL2,FAS,NOTCH1,PRKCQ,STAT5B,TCF7*	7
	synthesis of sphingolipid	UP	5.01E-04	*ADCYAP1,B3GALT4,B4GALT1,B4GALT6,ELOVL1,FAS*	6
	transmembrane potential of mitochondria	UP	2.53E-04	*APAF1,ARID1A,BCL2,CD47,FAS,LCK,MAP2K4,MYCN,PRF1,TRIB2,ZBTB16*	11
	turnover of phosphatidylinositol	UP	8.34E-04	*ADCYAP1,CD247,CD3D,CD3G*	4
	tyrosine phosphorylation of protein	UP	5.73E-04	*ABL2,CBL,CD247,CD47,CD7,DUSP1,EPHB6,FAS,LCK,SH2D1A,ZAP70*	11
	ubiquitination of protein	UP	9.65E-05	*BCL2,CAND1,CBL,FBXW2,KLHL9,MAPK15,PPIL2,RASSF1,SKP2,SMURF2,SOCS7,TSC1,UBR1,UBR2*	14

Downregulated genes	DNA damage checkpoint of cells	DOWN	5.96E-04	*CEBPA,CHEK1,PRKCD*	3
	activation of dendritic cells	DOWN	6.73E-04	*CD40,HMGB1,HSP90B1,IL18,LTBR,TLR2,TLR4,TNFRSF1B*	8
	activation of phagocytes	DOWN	9.31E-05	*ADAM9,CD40,CSF1R,ELANE,FPR1,HMGB1,HSP90B1,IL18,KLF4,LTBR,NDRG1,PRG2,PRTN3,PTGS2,S100A10,TLR2,TLR4,TNFRSF1B,TREM*	19
	activation of polysaccharide	DOWN	9.36E-04	*HMGB1,HTATIP2,TLR4*	3
	adhesion of eukaryotic cells	DOWN	4.52E-04	*ADAM15,ADAM9,ADRB2,ANXA5,APLP2,CDH2,CTNNA1,CTS2,CYTIP,E2F1,ELANE,FUT4,FYB,GFI1,HBEGF,HMGB1,IFNGR1,IL18,JAG1,L*	29
	apoptosis	DOWN	4.87E-07	*ABCG1,ADAM15,ADRB2,ANXA5,ARHGEF12,ATP1A1,BCL2L11,BCL3,BEX2,BHLHE40,BRF1,CAST,CD40,CD74,CDH2,CDK2,CEBPA,CEBPD*	104
	apoptosis of acute lymphoblastic leukemia cells	DOWN	5.96E-04	*BCL2L11,CD40,PRKCD*	3
	apoptosis of antigen presenting cells	DOWN	6.67E-05	*ABCG1,CD40,HMGB1,IL15RA,IL18,MCL1,NFE2L2,PLA2G4A,TLR2,TLR4,TNFRSF1B*	11
	apoptosis of blood cells	DOWN	1.10E-06	*ABCG1,BCL2L11,BCL3,CAST,CD40,CDK2,CHEK1,E2F1,GFI1,HMGB1,IFNGR1,IL15RA,IL18,IRAK3,IRS2,KLF4,LTBR,MCL1,MXD1,NFATC2,I*	30
	apoptosis of epithelial cells	DOWN	2.94E-08	*ABCG1,BCL2L11,CD40,CDH2,CEBPA,E2F1,ID1,IL15RA,IL18,IRAK3,MCL1,MYLK,NFE2L2,PLA2G4A,PLAUR,PRKCD,PTGS2,SKI,TCF4,TLR2,*	21
	apoptosis of eukaryotic cells	DOWN	2.36E-05	*ABCG1,ADRB2,ATP1A1,BCL2L11,BCL3,BEX2,BHLHE40,CAST,CD40,CDH2,CDK2,CEBPA,CEBPD,CHEK1,CSF1R,CSNK2A2,CTBP2,CTSB,DE*	81
	apoptosis of leukocytes	DOWN	7.56E-07	*ABCG1,BCL2L11,BCL3,CAST,CD40,CDK2,CHEK1,E2F1,GFI1,HMGB1,IFNGR1,IL15RA,IL18,IRAK3,IRS2,KLF4,LTBR,MCL1,NFATC2,NFE2L2,*	29
	apoptosis of macrophages	DOWN	9.10E-05	*ABCG1,HMGB1,IL18,MCL1,NFE2L2,PLA2G4A,TLR2,TLR4,TNFRSF1B*	9
	apoptosis of normal cells	DOWN	4.38E-06	*ABCG1,ADRB2,ATP1A1,BCL2L11,BCL3,CAST,CD40,CDH2,CDK2,CEBPA,CHEK1,CSF1R,CSNK2A2,CTBP2,CTSB,E2F1,FOXO1,GFI1,GNAS,I*	56
	apoptosis of peritoneal macrophages	DOWN	3.06E-05	*ABCG1,MCL1,NFE2L2,PLA2G4A,TLR4*	5
	apoptosis of phagocytes	DOWN	3.86E-06	*ABCG1,CAST,CD40,HMGB1,IL15RA,IL18,LTBR,MCL1,NFE2L2,PLA2G4A,PRKCD,PRTN3,TLR2,TLR4,TNFRSF1B,TREM1*	16
	apoptosis of tumor cells	DOWN	2.71E-04	*BCL2L11,CD40,CDK2,CTSB,E2F1,FOXO1,HBEGF,JAG1,MCL1,NFE2L2,NR4A2,PLAUR,PRKCD,RHOB,TNFSF13B,TRIB1*	16
	cell death	DOWN	2.11E-07	*ABCG1,ADAM15,ADRB2,ALDH3B1,ANXA5,ARHGEF12,ATP1A1,ATP2B1,BAG5,BCL2L11,BCL3,BEX2,BHLHE40,BRF1,CAST,CD40,CD74,C*	119
	cell death of blood cells	DOWN	4.89E-07	*ABCG1,BCL2L11,BCL3,CAST,CD40,CDK2,CHEK1,CST3,CTSB,E2F1,ELANE,GFI1,HMGB1,IFNGR1,IL15RA,IL18,IRAK3,IRS2,KLF4,LTBR,MCL*	33
	cell death of epithelial cells	DOWN	3.09E-08	*ABCG1,BCL2L11,CD40,CDH2,CEBPA,E2F1,ID1,IL15RA,IL18,IRAK3,MCL1,MYLK,NFE2L2,PLA2G4A,PLAUR,PRG2,PRKCD,PTGS2,SKI,TCF4*	22
	cell death of eukaryotic cells	DOWN	5.09E-06	*ABCG1,ADRB2,ALDH3B1,ATP1A1,ATP2B1,BAG5,BCL2L11,BCL3,BEX2,BHLHE40,CAST,CD40,CDH2,CDK2,CEBPA,CEBPD,CHEK1,CSF1R,C*	96
	cell death of leukocytes	DOWN	2.98E-07	*ABCG1,BCL2L11,BCL3,CAST,CD40,CDK2,CHEK1,CST3,CTSB,E2F1,ELANE,GFI1,HMGB1,IFNGR1,IL15RA,IL18,IRAK3,IRS2,KLF4,LTBR,MCL*	32
	cell death of macrophages	DOWN	7.27E-05	*ABCG1,CD40,HMGB1,IL18,MCL1,NFE2L2,PLA2G4A,TLR2,TLR4,TNFRSF1B*	10
	cell death of myeloid cells	DOWN	6.22E-04	*CAST,CDK2,ELANE,IRAK3,LTBR,MCL1,PRKCD,PRTN3,TREM1*	9
	cell death of normal cells	DOWN	6.37E-06	*ABCG1,ADRB2,ATP1A1,BAG5,BCL2L11,BCL3,CAST,CD40,CDH2,CDK2,CEBPA,CHEK1,CSF1R,CSNK2A2,CST3,CTBP2,CTSB,E2F1,ELANE,F*	64
	cell division process of B lymphocytes	DOWN	1.39E-04	*BCL2L11,CD40,E2F1,KLF4,KLF9,TNFSF13B*	6
	cell division process of eukaryotic cells	DOWN	4.78E-05	*ASCL2,BCL2L11,BCL3,BHLHE40,CAST,CD40,CDK14,CDK2,CEBPA,CEBPD,CHEK1,CHFR,CREG1,CSF1R,CSNK2A2,DEGS1,E2F1,ELAVL1,ERI*	47
	cell division process of lymphocytes	DOWN	4.07E-04	*BCL2L11,CD40,CDK2,E2F1,GFI1,KLF4,KLF9,NFATC2,TNFSF13B*	9
	cell movement of bone marrow cells	DOWN	8.57E-04	*CAST,CTSB,ELANE,FPR1,FUT4,HMGB1,IL18,NFE2L2,PRTN3,PTGS2,SIRPA,TLR2,TLR4,TNFRSF1B*	14
	cell stage of eukaryotic cells	DOWN	2.14E-04	*ASCL2,BCL2L11,BCL3,BHLHE40,CD40,CDK2,CEBPA,CEBPD,CHEK1,CHFR,CREG1,DEGS1,E2F1,ELAVL1,GFI1,GRN,ID1,JUNB,KLF4,LDLRAI*	34
	cell stage of lymphocytes	DOWN	2.18E-04	*BCL2L11,CD40,CDK2,E2F1,GFI1,KLF4,TNFRSF1B*	7
	development of lymphocytes	DOWN	3.61E-05	*BCL2L11,BCL3,BST1,CD40,CD74,CDK2,CEBPA,CHEK1,CST3,E2F1,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA,IL18,JUNB,L*	36
	developmental process of mononuclear leukocytes	DOWN	2.35E-05	*BCL2L11,BCL3,BST1,CD40,CD74,CDK2,CEBPA,CHEK1,CSF1R,CST3,E2F1,FOXO1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,IL13RA1,IL15*	41
	developmental process of osteoclasts	DOWN	2.95E-04	*CSF1R,IFNGR1,IL18,IRAK3,JAG1,JUNB,MITF,NFATC2,TLR2,TLR4,ZBTB7A*	11
	differentiation of blood cells	DOWN	1.98E-04	*BCL3,CD40,CD74,CEBPA,CEBPD,CSF1R,E2F1,FZD8,GFI1,HSP90B1,ID1,IFNGR1,IL13RA1,IL15RA,IL18,JAG1,JUNB,KLF4,LTBR,MBD2,MY*	33
	differentiation of cells	DOWN	6.49E-09	*ADAM22,ADRB2,ARL4A,ASCL2,BCL2L11,BCL3,BEX1,BEX2,BHLHE40,CAST,CD40,CD74,CDH2,CDK2,CEBPA,CEBPD,CREG1,CSDA,CSF1R,*	86
	differentiation of connective tissue cells	DOWN	3.04E-04	*ADRB2,ARL4A,CAST,CEBPA,CEBPD,CSF1R,FOSL2,FOXO1,GRN,IL18,IRAK3,IRS2,JAG1,JUNB,KLF4,MITF,NFATC2,NFIC,PLAUR,SKI,TLR2*	24
	differentiation of mononuclear leukocytes	DOWN	1.20E-04	*BCL3,CD40,CD74,CEBPA,CSF1R,FZD8,GFI1,HSP90B1,ID1,IFNGR1,IL13RA1,IL15RA,IL18,JUNB,KLF4,MBD2,MYLK,NDFIP1,NFATC2,PRK*	27
	differentiation of osteoclasts	DOWN	3.40E-04	*CSF1R,IL18,IRAK3,JAG1,JUNB,MITF,NFATC2,TLR2,TLR4,ZBTB7A*	10
	entry into cell division process of B lymphocytes	DOWN	1.77E-04	*CD40,KLF4,TNFSF13B*	3
	entry into cell division process of lymphocytes	DOWN	1.42E-04	*CD40,E2F1,GFI1,KLF4,TNFSF13B*	5
	entry into cell division process of normal cells	DOWN	8.82E-04	*CD40,CDK2,E2F1,GFI1,JUNB,KLF4,PRKCD,TNFSF13B*	8
	entry into S phase of lymphocytes	DOWN	5.50E-05	*CD40,E2F1,GFI1,KLF4*	4
	entry into S phase of normal cells	DOWN	3.67E-04	*CD40,CDK2,E2F1,GFI1,JUNB,KLF4,PRKCD*	7
	expression of DNA	DOWN	8.89E-04	*ASCL2,BCL3,BHLHE40,CEBPA,CREG1,CTBP2,DNMT1,DVL2,E2F1,ENY2,FOSL2,FOXO1,GMEB1,HMGB1,ID1,JUNB,KLF4,KLF9,MITF,MXD*	36
	formation of tumor	DOWN	2.77E-04	*CDK2,CHFR,DACH1,E2F1,HBEGF,HMGB1,IL18,LTBR,MCL1,NFE2L2,PCYT1B,PLAUR,PRKCD,PTGS2,RRM1,TNFRSF1B,TOB1*	17
	function of leukocytes	DOWN	9.21E-04	*CSF1R,CTSZ,GFI1,HMGB1,IL18,TLR2,TLR4,TNFSF13B,TOB1*	9
	growth of eukaryotic cells	DOWN	5.99E-05	*BCL2L11,CAST,CD40,CDK2,CEBPA,CEBPD,CHEK1,CHFR,CREG1,CSF1R,DACH1,DEGS1,E2F1,ELANE,EMILIN2,EREG,FOXO1,GFI1,GFM1,*	58
	hatching	DOWN	5.96E-04	*GRN,MCL1,PTGS2*	3
	implantation	DOWN	3.30E-04	*GRN,HBEGF,KLF9,MCL1,PLA2G4A,PTGS2*	6
	infiltration of leukocytes	DOWN	6.44E-04	*ADRB2,CD40,CSF1R,CTSB,ELANE,FUT4,HMGB1,IL18,LTBR,MYLK,NFATC2,NFE2L2,PRKCD,PRTN3,PTGS2,S100A10,TLR2,TLR4,TNFRSF1*	19
	influx of neutrophils	DOWN	2.01E-04	*HMGB1,IL18,IRAK3,LTBR,PLAUR,TLR2*	6
	interphase of eukaryotic cells	DOWN	6.02E-04	*ASCL2,BCL2L11,BCL3,BHLHE40,CD40,CDK2,CEBPA,CEBPD,CHEK1,CREG1,DEGS1,E2F1,ELAVL1,GFI1,ID1,JUNB,KLF4,LMNA,MXD1,NAS*	26
	invasion of cells	DOWN	3.97E-06	*ADAM9,CAPN2,CDH2,CHFR,CSNK2A2,CTBP2,CTSB,GRN,HBEGF,HDLBP,HMGB1,HTATIP2,ID1,IL18,IRS2,JUNB,KLF4,LDLRAP1,MBD2,A*	35
	invasion of eukaryotic cells	DOWN	6.99E-06	*ADAM9,CAPN2,CDH2,CHFR,CSNK2A2,CTBP2,CTSB,GRN,HBEGF,HDLBP,HMGB1,HTATIP2,ID1,IL18,IRS2,JUNB,KLF4,LDLRAP1,MBD2,A*	33
	Lymphocyte homeostasis	DOWN	7.91E-05	*BCL2L11,BCL3,CD40,CD74,CDK2,CEBPA,CHEK1,CST3,E2F1,FZD8,GFI1,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA,IL18,JUNB,LTBR,MBD2,M*	34
	migration of cells	DOWN	5.46E-04	*ADAM15,ADAM9,ADRB2,AMOTL1,APLP2,CAPN2,CD40,CD74,CDH2,CSF1R,CTBP2,CTSB,CTS2,CYTIP,ELANE,EREG,FHL1,FOXO1,FPR1,*	58
	migration of eukaryotic cells	DOWN	4.31E-04	*ADAM15,ADAM9,ADRB2,AMOTL1,APLP2,CAPN2,CD40,CD74,CDH2,CSF1R,CTBP2,CTSB,CTS2,CYTIP,ELANE,EREG,FHL1,FOXO1,FPR1,*	55
	proliferation of B lymphocytes	DOWN	5.08E-04	*BCL2L11,CD40,CD74,IL13RA1,IL18,IRS2,JUNB,KLF4,KLF9,NFATC2,PELI1,PRKCD,TLR2,TLR4,TNFSF13B*	15
	proliferation of cells	DOWN	1.34E-07	*ABCG1,ADAM15,ADRB2,ASCL2,ATP2B1,BCL2L11,BCL3,BEX1,BEX2,BRF1,CAPN2,CD40,CD74,CDH2,CDK14,CDK2,CEBPA,CEBPD,CGREF*	108
	proliferation of eukaryotic cells	DOWN	7.91E-08	*ABCG1,ADAM15,ADRB2,ASCL2,ATP2B1,BCL2L11,BEX1,BEX2,BRF1,CAPN2,CD40,CD74,CDH2,CDK14,CDK2,CEBPA,CEBPD,CHEK1,CSD*	92
	proliferation of mononuclear leukocytes	DOWN	2.47E-05	*ABCG1,ADRB2,BCL2L11,CD40,CD74,CDK2,CSF1R,CTSZ,FYB,GFI1,HMGB1,HSP90B1,IFNGR1,IL13RA1,IL15RA,IL18,IRS2,JAG1,JUNB,KLF*	34
	proliferation of normal cells	DOWN	7.80E-07	*ABCG1,ADAM15,ADRB2,ASCL2,ATP2B1,BCL2L11,BEX1,CAPN2,CD40,CD74,CDH2,CDK2,CEBPA,CEBPD,CHEK1,CSF1R,CTNNA1,CTS2,D*	70
	proliferation of smooth muscle cells	DOWN	2.41E-06	*ATP2B1,CAPN2,CEBPD,ELANE,ELAVL1,EREG,FHL1,FOXO1,GNA11,HBEGF,HMGB1,KLF4,NFE2L2,PLAUR,PTGS2,TLR2,TLR4,TRIB1*	18
	proliferation of vascular smooth muscle cells	DOWN	2.34E-04	*CEBPD,ELAVL1,FHL1,GNA11,NFE2L2,PLAUR,TLR2,TLR4,TRIB1*	9
	quantity of antigen presenting cells	DOWN	3.78E-04	*BCL2L11,CEBPA,CSF1R,HBEGF,IFNGR1,IL15RA,IL18,LTBR,PDGFC,PRKCD,SIRPA,TLR2,TLR4*	13
	quantity of granulocytes	DOWN	2.84E-04	*BCL2L11,CEBPA,HMGB1,IL18,LTBR,PRKCD,PROK2,PRTN3,RAG1,TLR2,TLR4,TNFRSF1B*	12
	quantity of myeloid cells	DOWN	2.62E-04	*BCL2L11,CEBPA,HMGB1,IL18,LTBR,MXD1,PRKCD,PROK2,PRTN3,RAG1,TLR2,TLR4,TNFRSF1B*	13
	quantity of phagocytes	DOWN	2.77E-04	*BCL2L11,CEBPA,CSF1R,HBEGF,HMGB1,IFNGR1,IL15RA,IL18,LTBR,PDGFC,PRKCD,PROK2,PRTN3,SIRPA,TLR2,TLR4,TNFRSF1B*	17
	reproductive process of blastocyst	DOWN	1.38E-04	*GRN,HBEGF,MCL1,PTGS2*	4
	S phase of B lymphocytes	DOWN	6.97E-06	*BCL2L11,CD40,E2F1,KLF4*	4
	S phase of lymphocytes	DOWN	3.21E-06	*BCL2L11,CD40,CDK2,E2F1,GFI1,KLF4*	6
	survival of cells	DOWN	4.05E-04	*ABCC6,BCL2L11,BCL3,BEX2,BHLHE40,CD40,CD74,CDK2,CEBPD,CHEK1,CHFR,CSF1R,DNMT1,E2F1,EHD4,EMILIN2,FOXO1,HBEGF,HMG*	45
	survival of eukaryotic cells	DOWN	3.10E-04	*BCL2L11,BCL3,BEX2,BHLHE40,CD40,CD74,CDK2,CEBPD,CHEK1,CHFR,CSF1R,DNMT1,E2F1,EHD4,EMILIN2,FOXO1,HBEGF,HMGB1,HTA*	42
	survival of mammalia	DOWN	4.56E-04	*ADRB2,B3GALNT1,BCL2L11,CD40,CD74,CEBPA,CTSB,DNMT1,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA,IL18,LTBR,MME,MSRA,NFE2L2,P*	26
	survival of normal cells	DOWN	3.18E-04	*BCL2L11,BCL3,CD40,CEBPD,CHEK1,CSF1R,DNMT1,E2F1,EHD4,FOXO1,HMGB1,IL18,IRS2,MCL1,MITF,MVP,NDRG1,PPP3CA,PROK2,PS*	27
	survival of rodents	DOWN	5.06E-04	*ADRB2,B3GALNT1,BCL2L11,CD40,CD74,CEBPA,CTSB,DNMT1,HMGB1,HSP90B1,ID1,IFNGR1,IL15RA,IL18,LTBR,MSRA,NFE2L2,PTGS2,F*	25
	synthesis of carbohydrate	DOWN	4.23E-04	*B3GALNT1,CD40,CEBPA,CSF1R,HS6ST1,IL18,IRS2,MAN2B1,PCYT1B,PLA2G4A,PLD1,PLEK,PPP1R3D,PTGS2,SCAP,SLC35D1,VAC14*	17
	synthesis of prostaglandin	DOWN	9.02E-04	*CD40,CD74,MITF,NFATC2,PLA2G4A,PTGS2,VIM*	7
	synthesis of prostaglandin D2	DOWN	3.48E-04	*MITF,PLA2G4A,PTGS2*	3
	transcription	DOWN	1.96E-04	*ADRB2,ASCL2,BASP1,BCL3,BHLHE40,BRF1,CD40,CDK2,CEBPA,CEBPD,CREG1,CSDA,CTBP2,DACH1,DNMT1,DVL2,E2F1,ENY2,EREG,FO.*	75

Table S1. Enriched pathways for the 286 upregulated and 403 downregulated genes of CEBPAsilenced cases. Ingenuity pathway analysis is performed on the 286 upregulated and 403 downregulated genes. The enriched functional pathways (P<0.001 and > 3 genes per pathway) are sorted on their presence in: enriched in both upregulated and downregulated genes, enriched in solely the upregulated genes and enriched in solely the downregulated genes. Upregulation and downregulation is relative for the *CEBPA*^{silenced} group compared to normal cells (CD34+ group).

Canonical Pathways	Regulation	-Log10(p-value)	Gene symbols
T Cell Receptor Signaling	upregulated	6.92	*MAP2K4, CD247, CD3G, LCK, PRKCQ, GRAP2, SOS1, ZAP70, NFATC4, CD3D, ATM*
	downregulated	-	-
PKCθ Signaling in T Lymphocytes	upregulated	5.40	*MAP2K4, CD247, CD3G, LCK, PRKCQ, GRAP2, SOS1, ZAP70, NFATC4, CD3D, ATM*
	downregulated	-	-
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	upregulated	5.04	*CD247, CD3G, PRF1, APAF1, CD3D, FAS, BCL2*
	downregulated	-	-
iCOS-iCOSL Signaling in T Helper Cells	upregulated	4.96	*CD247, CD3G, LCK, PRKCQ, GRAP2, ZAP70, TRAT1, NFATC4, CD3D, ATM*
	downregulated	0.32	*CD40, NFATC2, PPP3CA*
Type I Diabetes Mellitus Signaling	upregulated	4.77	*MAP2K4, CD247, CD3G, PRF1, APAF1, JAK2, SOCS7, CD3D, FAS, BCL2*
	downregulated	-	-
CTLA4 Signaling in Cytotoxic T Lymphocytes	upregulated	4.65	*CD247, CD3G, LCK, GRAP2, ZAP70, TRAT1, JAK2, CD3D, ATM*
	downregulated	-	-
CD28 Signaling in T Helper Cells	upregulated	4.59	*MAP2K4, CD247, CD3G, LCK, PRKCQ, GRAP2, ZAP70, NFATC4, CD3D, ATM*
	downregulated	-	-
Prolactin Signaling	upregulated	3.46	*PRKCQ, SOS1, JAK2, SOCS7, STAT5B, TCF7, ATM*
	downregulated	0.24	*PRKCI, PRKCD*
Calcium-induced T Lymphocyte Apoptosis	upregulated	3.38	*CD247, CD3G, LCK, PRKCQ, ZAP70, CD3D*
	downregulated	1.86	*PRKCI, PRKCD, NFATC2, CAPN2, PPP3CA*
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	upregulated	3.35	*MAP2K4, CD247, CD3G, SOS1, ZAP70, NFATC4, CD3D*
	downregulated	0.49	*TOB1, NFATC2, PPP3CA*
Myc Mediated Apoptosis Signaling	upregulated	3.21	*MAP2K4, SOS1, APAF1, FAS, ATM, BCL2*
	downregulated	-	-
Role of NFAT in Regulation of the Immune Response	upregulated	3.18	*CD247, CD3G, LCK, PRKCQ, SOS1, ZAP70, NFATC4, CD3D, ATM, ATF2*
	downregulated	0.61	*GNAS, CSNK1G3, GNA11, NFATC2, GNG7, PPP3CA*
Glycosphingolipid Biosynthesis - Globoseries	upregulated	-	-
	downregulated	3.74	*NAGA, GLA, HEXB, B3GALNT1, FUT4*

Table S2. Enriched canonical pathways for the 286 upregulated and 403 downregulated genes. Ingenuity pathway analysis is performed on the 286 upregulated and 403 downregulated genes. The enriched canonical pathways (P<0.001) are sorted on their presence in the upregulated and downregulated genes, Upregulation and downregulation is relative for the *CEBPA*^{silenced} group compared to normal cells (CD34+ group).

Regulation	Matrix name	Recognized factors	Fold increase	*P*-value
Upregulated	V$DMRT1_01	DMRT1	2.2661	0.00023791
Upregulated	V$ARNT_01	arnt, arnt-L	2.2034	4.1679E-05
Upregulated	V$CETS1P54_03	Ets-1, Ets-1 deltaVII, c-Ets-1, c-Ets-	1.9741	0.00087532
Upregulated	V$SZF11_01	SZF1-isoform3	1.9113	0.00082443
Upregulated	V$ELK1_02	Elk-1, Elk1-isoform1	1.8956	0.00093331
Upregulated	V$NRF1_Q6	NRF-1	1.8121	0.00017664
Upregulated	V$E2F6_01	E2F-6	1.7668	0.00051006
Upregulated	V$R_01	R	1.697	0.00037116
Upregulated	V$USF_02	USF, usf1	1.5658	1.6743E-07
Upregulated	V$MAX_01	max-isoform2	1.5469	3.2868E-07
Upregulated	V$AHRARNT_02	AhR, AhR2, arnt, arnt-L	1.5348	0.00090063
Upregulated	V$WHN_B	FOXN1	1.5285	0.00013026
Upregulated	V$USF_Q6	USF, USF2, USF2A-delta-H, USF2a,	1.526	4.6985E-05
Downregulated	V$ARNT_01	arnt, arnt-L	1.7106	0.00070203
Downregulated	V$CMYC_02	c-Myc, c-Myc I	1.6331	0.00082238
Downregulated	V$ARNT_02	arnt, arnt-L	1.5848	4.5006E-05
Downregulated	V$CLOCKBMAL_Q6	Clock:BMAL1, Clock:BMAL2	1.5345	0.00079547

Table S3. Transcription factor binding sites for the upregulated (n=286) and downregulated (n=403) genes specific for the *CEBPA*^{silenced} leukemias. Transcription factors binding sites (TFBS) that are enriched among the promoter regions (2Kb upstream) of the 286 upregulated and 403 downregulated genes (HG19). The 2Kb upstream sequences for each

gene is gathered from the UCSC database. Regulation: genes that are upregulated and downregulated compared to CD34+. Matrix name: Enriched transcription factor binding name. Recognized factors: Binding sites that are recognized. Fold-increase: the ratio that a TFBS is enriched compared to 5000 randomly selected genes.

<Table S4 is not included>

Table S4. C/EBPα target genes. Gene symbols: Genes that were closest to the detected regions-of-interest. Chromosome: chromosome number. Genomic location start -stop: The genomic location for which the binding was detected.

Transcription Factor	Recognized factors	Fold increase	*P*-value	Site
CEBP_Q2	C/EBP, C/EBPalpha, C/EBPalpha(p20),	4.2672	2.5008E-38	consensus
CEBP_Q2_01	C/EBPalpha, C/EBPalpha(p20, p30), C/EBPbeta(LAP, p20, p34, p35), C/EBPgamma, CRP2, CRP3, LAP*-NF-M, NF-IL6-1, NF-IL6-2, NF-IL6-3, cebpe	4.2866	2.1402E-39	consensus
CEBPA_01	C/EBP, C/EBPalpha, C/EBPalpha(p20), C/EBPalpha(p30)	3.9335	0	consensus
CEBPB_01	C/EBPbeta(LAP), C/EBPbeta(p35), CRP2	4.9584	1.0174E-39	consensus
CEBPB_02	C/EBPbeta(LAP), C/EBPbeta(p35), CRP2	5.262	5.8184E-42	consensus
CEBPDELTA_Q6	CRP3	3.8018	3.7812E-29	-

Table S5. Enriched transcription factor binding sites for the detected binding regions. TFBS analyses among the detected binding regions (based on wild-type C/EBPα from chip-on-ChIP), resulted in the detection of the C/EBPα consensus sites.

Regulation	Function	Function Annotation	*P*-value	Genes	Nr. of genes
Upregulated	Cell-mediated Immune Response	apoptosis of T lymphocytes	4.50E-04	BCL2, CASP1, CD47, MAP2K4	4
		T cell development	2.64E-03	BCL2, CASP1, CCR9, CD47, MAP2K4	5
	Hematological System Development and Function	apoptosis of T lymphocytes	4.50E-04	BCL2, CASP1, CD47, MAP2K4	4
		hematopoiesis	2.22E-03	BCL2, CASP1, CCR9, CD47, MAP2K4, OGT	6
		T cell development	2.64E-03	BCL2, CASP1, CCR9, CD47, MAP2K4	5
		cell movement of neutrophils	5.79E-03	B4GALT1, CASP1, CD47	3
		proliferation of T lymphocytes	6.55E-03	B3GNT2, BCL2, CD47, MAP2K4	4
	Hematopoiesis	apoptosis of T lymphocytes	4.50E-04	BCL2, CASP1, CD47, MAP2K4	4
		hematopoiesis	2.22E-03	BCL2, CASP1, CCR9, CD47, MAP2K4, OGT	6
		T cell development	2.64E-03	BCL2, CASP1, CCR9, CD47, MAP2K4	5
		cell movement of bone marrow cells	3.91E-03	B4GALT1, CASP1, CD47	3
	Cell Death	cell death of leukemia cells	6.52E-05	BCL2, CASP1, CD47	3
		apoptosis of B lymphocytes	3.73E-04	BCL2, CASP1, CD47	3
		killing of eukaryotic cells	4.11E-04	BCL2, CD47, CEBPG	3
		apoptosis of T lymphocytes	4.50E-04	BCL2, CASP1, CD47, MAP2K4	4
		cell viability of normal cells	4.96E-04	BCL2, CD47, OGT	3
		apoptosis of cardiomyocytes	8.15E-04	BCL2, CASP1, MAP2K4	3
		cell viability of eukaryotic cells	8.50E-04	BCL2, CASP1, CD47, OGT	4
		apoptosis of fibroblasts	3.09E-03	BCL2, CASP1, MAP2K4	3
	Cell Morphology	autophagy of eukaryotic cells	1.14E-03	BCL2, CASP1, GOPC	3
		transmembrane potential of mitochondria	2.47E-03	BCL2, CD47, MAP2K4	3
	Cellular Development	apoptosis of T lymphocytes	4.50E-04	BCL2, CASP1, CD47, MAP2K4	4
		developmental process of leukocytes	2.54E-03	BCL2, CASP1, CCR9, CD47, MAP2K4, OGT	6
		T cell development	2.64E-03	BCL2, CASP1, CCR9, CD47, MAP2K4	5
	Cellular Function and Maintenance	apoptosis of T lymphocytes	4.50E-04	BCL2, CASP1, CD47, MAP2K4	4
		autophagy of eukaryotic cells	1.14E-03	BCL2, CASP1, GOPC	3
		T cell development	2.64E-03	BCL2, CASP1, CCR9, CD47, MAP2K4	5
	Post-Translational Modification	modification of protein	8.30E-07	B3GNT2, B4GALT1, BCL2, CASP1, CD47, FBXW2, MAP2K4, NMT1, OGT, SEPX1, STK38	11
		moiety attachment of protein	8.02E-06	B3GNT2, B4GALT1, BCL2, CD47, FBXW2, MAP2K4, NMT1, OGT, STK38	9
		glycosylation of protein	2.15E-04	B3GNT2, B4GALT1, OGT	3
Downregulated	Cell death	killing of cells	1.64E-03	HSP90B1, MSRA, PLA2G4A	3
	Lipid metabolism	accumulation of lipid	6.42E-03	ACSL1, COL4A3BP, PLA2G4A	3
	Small Molecule Biochemistry	accumulation of lipid	6.42E-03	ACSL1, COL4A3BP, PLA2G4A	3
	Molecular Transport	accumulation of lipid	6.42E-03	ACSL1, COL4A3BP, PLA2G4A	3
	Connective Tissue Development and Function	differentiation of osteocytes	6.65E-03	FOSL2,IRAK3,TOB1	3
	Cellular Development	differentiation of osteocytes	6.65E-03	FOSL2,IRAK3,TOB1	3

Table S6. Enriched pathways for the 24 upregulated and 25 downregulated genes of *CEBPA*^{silenced} cases. Ingenuity pathway analysis is performed on the 24 upregulated and 25 downregulated genes. Enriched functional pathways are

considered if P<0.01 and the pathway contains more than 3 genes. Upregulation and downregulation is relative for the *CEBPA*^{silenced} group compared to normal cells (CD34+ group).

Transcription Factor	Recognized factors	Fold increase	P-value	SLC7A11	MRPL41	CEPT1	BCL2	UBTD1	ATP6V1B2	STK38	GOPC	CASP1	OGT	MAP2K4	CD47	ACAD8	YPEL1	CCR9	SSBP2	B3GNT2	HMGCR	FBXW2	NIN	SEPX1	B4GALT1	CEBPG	HMT1
V$AFP1_Q6	AFP1, ATBF1	3.8311	0.00080592	0	0	1	0	0	0	3	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0
V$EVI1_04	Evi-1	3.3417	0.00067471	2	0	3	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
V$SP1_Q6	Sp1, Sp1-isoform1, Sp1-isoform2	3.1249	5.6024E-06	0	8	0	0	2	0	0	0	0	0	0	1	0	3	0	0	3	0	0	0	3	2	0	0
V$GC_01		3.0977	1.0233E-05	0	8	0	0	2	0	0	0	0	0	0	1	0	2	0	0	3	0	0	0	3	2	0	0
V$SP4_Q5	sp4	2.8884	0.00023466	0	8	0	0	1	0	0	0	0	0	0	1	0	3	0	0	1	0	0	0	1	1	0	0
V$SP1_Q6_01	Sp1, Sp1-isoform1, Sp1-isoform2, sp3, sp3-isoform1, sp3-	2.8384	2.3341E-05	0	8	0	0	2	0	0	0	0	0	0	1	0	3	0	0	3	0	0	0	3	2	0	0
V$AP2ALPHA_03	AP-2alphaA	2.7243	0.00043649	0	2	2	0	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	2	0	4
V$RPS8_01	RPS8	2.7157	0.001	0	0	1	0	0	1	1	0	0	1	1	0	1	0	1	1	1	0	0	0	1	2	1	0
V$EVI1_01	Evi-1	2.5133	0.001	1	1	2	0	0	2	2	0	0	0	0	0	0	1	0	1	0	0	1	1	1	0	0	1
V$SP1_Q4_01	Sp1, Sp1-isoform1, Sp1-isoform2, Sp2, sp3, sp3-isoform1,	2.4267	0.00011068	0	8	1	0	2	0	0	0	0	0	0	1	0	3	0	0	4	0	0	0	3	2	0	0
V$CNOT3_01	CNOT3	2.1077	0.001	0	3	0	7	0	0	0	0	0	0	0	0	0	3	0	1	0	0	2	2	0	4	1	0
V$MUSCLE_INI_B		2.0079	0.00051498	0	11	0	5	1	0	1	0	0	0	1	2	0	1	0	3	1	0	0	1	1	1	0	0
V$MYCMAX_B	c-Myc, c-Myc l, max-isoform1, max-isoform2	1.7818	0.00041057	0	4	1	5	4	0	2	2	0	0	0	4	1	4	0	1	2	1	4	1	3	1	1	1
V$DMRT1_01	DMRT1	1.7216	4.3474E-05	5	0	4	0	0	2	2	2	5	4	2	4	2	7	4	1	4	0	0	4	1	0	3	4
V$AP4_Q6	AP-4, TFAP4	1.6665	0.00088342	0	24	2	0	0	0	0	0	1	3	2	2	0	0	1	0	3	0	0	0	3	2	1	0
V$AP2ALPHA_02	AP-2alphaA	1.6245	0.0008359	0	4	4	1	6	1	4	0	1	1	0	2	3	0	0	4	1	1	1	0	7	2	1	6
V$CEBPB_01	C/EBPbeta(LAP), C/EBPbeta(p35), CRP2	1.6187	0.00065267	4	0	3	1	3	5	2	5	2	2	3	0	1	1	0	1	2	4	3	3	2	1	1	1
V$PU1_Q4	PU.1-xbb1, SPI1	1.5834	1.6533E-05	5	2	4	2	1	8	22	2	5	0	1	2	4	1	6	5	1	3	2	2	2	7	2	3
V$E4BP4_01	E4BP4	1.5249	0.00023745	16	1	2	1	8	4	3	5	5	3	3	0	3	2	2	2	1	5	4	5	0	1	0	0
V$NMYC_01	N-Myc	1.5041	0.00011056	0	7	5	14	6	4	4	5	0	3	0	8	5	5	3	3	5	1	5	2	2	1	1	4
V$BEN_01	Ben	1.4973	0.00021861	0	15	5	14	5	1	1	2	0	2	0	5	3	5	2	5	2	0	3	2	3	5	5	2
V$NRF1_Q6	NRF-1	1.4851	0.0007888	0	5	6	10	6	4	4	2	0	3	0	9	1	6	1	2	2	4	3	1	0	1	2	2
V$NKX61_01	NKX6A	1.4731	0.00003336	14	0	6	0	4	6	7	3	6	8	14	3	2	1	7	5	3	8	5	3	2	0	2	5
V$OCT1_07	Oct-1, POU2F1, POU2F1a, POU2F1b, POU2F1c	1.4559	0.00013524	10	0	8	1	4	8	2	6	6	7	5	0	8	0	7	1	2	6	4	5	3	1	1	1
V$CREBP1_01	ATF-2-xbb4	1.4489	0.00028742	9	1	3	8	3	8	5	8	10	0	4	0	4	2	2	4	2	8	4	4	0	2	2	1
V$CEBPGAMMA_Q6	C/EBPgamma	1.4181	0.00086097	14	0	6	2	4	6	2	7	6	6	1	4	4	2	2	3	1	4	2	4	0	1	1	3
V$DMRT3_01	DMRT3	1.4008	0.0005099	6	0	6	1	6	7	2	8	9	9	5	5	5	3	8	1	1	2	5	4	1	2	1	3
V$IPF1_Q6	PDX1, ipf1	1.3754	0.00010223	18	1	8	0	4	7	6	5	11	7	10	6	9	2	8	8	4	8	5	6	2	1	3	2
V$VBP_01	TEF-xbb1, Thyrotroph embryonic factor, Thyrotroph	1.3724	0.00019089	20	1	4	2	12	9	5	13	9	8	8	1	4	2	4	4	3	7	6	8	1	2	1	2
V$R_01	R	1.3354	0.00077965	1	27	5	18	7	1	3	0	1	2	4	4	3	7	2	6	8	4	3	3	10	4	6	2
V$SOX9_B1	Sox9	1.332	0.0007278	7	3	4	2	6	13	5	9	3	8	6	11	9	3	8	2	3	5	4	5	3	2	6	2
V$LHX3_01	LHX3a, LHX3b, Lhx3a, Lhx3b	1.3067	0.00066458	11	2	10	1	3	7	14	11	8	8	13	6	5	1	11	10	2	6	6	3	1	0	2	2
V$ARNT_01	arnt, arnt-L	1.3025	0.001	2	8	8	17	9	8	7	9	4	8	3	6	7	6	6	2	6	2	6	4	4	5	4	6
V$IPF1_Q4_01	Ipf1, ipf1, ipf1:Pbx	1.3017	0.00092139	16	0	9	0	3	7	5	5	13	8	7	9	14	3	5	8	4	7	5	6	2	2	3	2
V$E2F_Q2	DP-1, E2F, E2F-1, E2F-3, E2F-3A, E2F-3B, E2F-3a, E2F-4,	1.3016	0.00075182	0	13	8	24	9	1	3	3	0	0	5	7	2	17	2	7	12	1	2	5	10	14	12	3
V$S8_01	S8	1.2978	0.00084432	17	2	14	1	7	6	8	7	10	11	9	3	4	2	7	4	4	5	4	4	4	1	5	4

Table S7. Enriched TFBSs in the 2Kb upstream promoter regions of the 24 upregulated genes. Transcription factors binding sites (TFBS) that are enriched among the promoter regions (2Kb upstream) of the 24 upregulated genes (HG19) that also showed binding by *CEBPA* within the 32D-model-system. The 2Kb upstream sequences for each gene is gathered from the UCSC database. Transcription Factor: The enriched transcription factor name. Recognized factors: Binding sites that are recognized. Fold-increase: the ratio that a TFBS is enriched compared to 5000 randomly selected genes. *P-value*: Significance. The numbers under the gene-symbols indicate how often the representative TF was detected. An empty field illustrates that the TFBS was not enriched for the particular group.

CHAPTER

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

BLOOD

February 2011 | Volume 117 | Issue 8 | doi: 10.1182/blood-2010-09-307280

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

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.

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" [199,200].

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 [165,175]. Two proteins may be translated from the *CEBPA* transcripts, i.e., a 42kDa (p42) and a shorter 30kDa (p30) protein both translated from the same mRNA transcript. 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 (CN) AML (CN-AML) with an incidence of 5-14% [27,28,31,36-42]. Two main types of mutations can be distinguished: N-terminal frame-shift mutations resulting in the translation of a 30-kDa protein only, and the C-terminal in-frame mutations in the basic zipper region affecting DNA-binding and homo- and heterodimerization [28,43]. As a consequence, these mutations create an imbalance between proliferation and differentiation of hematopoietic progenitors [40,180].

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}) [35,45-48]. In the majority of *CEBPA*^{dm} AML, both alleles are mutated [46]. 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 [31,36,38,41]. More recent data suggest that this favorable outcome is mainly observed in AML with *CEBPA*^{dm} and not *CEBPA*^{sm} [35,45-48]. 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 [35]. 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[18,34].

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) [201-204], and 989 patients on AMLSG protocols AMLHD93 (n=74) [205], AML HD98A (n=313) [206], AMLSG 07-04 (n=376; ClinicalTrials.gov Identifier NCT00151242), AML SHG 02-95 (n=94) [207], 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, the 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 [24,208,209]. *CEBPA*^{dm} and *CEBPA*^{sm} AML were identified by denaturing high-performance liquid chromatography (dHPLC) or direct sequencing as described [35]. Cases that carried an insertion polymorphism [35,48] (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 [31,35,38].

Gene Expression Profiling

Data from GEP analysis were available in 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 [18,181]. 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).

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.

Outcome measures of the HOVON-SAKK and AMLSG-cohorts were comparable (log-rank test overall survival (OS), *P*= .08; event-free survival (EFS), *P*= .47; Figure S1A and S1B, respectively). There were no statistical differences in outcome in patients receiving autologous or allogeneic hematopoietic stem cell transplantation between the HOVON-SAKK and AMLSG-cohorts (log-rank test OS, *P*= .68; EFS, *P*= .89; Figure S2A and S2B respectively).

For univariate analysis, significance was assessed using the stratified log-rank test and Kaplan-Meier estimates for OS, EFS and relapse-free survival (RFS). The recommended consensus criteria [210] were used for the definition of CR and survival endpoints such as OS, EFS, and RFS. Multivariate analysis was performed by using stratified Cox's proportional hazard model. For all analyses, a *P-value* less or equal than .05 was considered statistically significant and for survival analyses, *P-value*s were computed using the full time span. Note that the close testing procedure [211] was applied and a correction for multiple testing [212] 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 (Supplementary material: creation and evaluation of the *CEBPA*^{dm} predictive signature).

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).

CEBPA germline mutation analysis

Five out of 71 (7%) *CEBPA-mutant* AML patients analyzed carried *CEBPA* germline mutations: in two of the 5 patients, the germline mutation was localized in the N-terminus and both acquired a C-terminal mutation. Both patients had a family history of AML and were diagnosed at young age. In the remaining three patients, the germline mutation was in the C-terminus; one of these patients gained an additional N-terminal mutation and the second patient an additional C-terminal mutation at the time of AML diagnosis. None of these three patients had a family history of AML. Alignment to distinct SNP databases (http://www.ncbi.nlm.nih.gov/sites/snp; http://genome.ucsc.edu/cgi-bin/hgGateway; http://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene) did not identify one of these germline sequence variations as a polymorphism. Using the PolyPhen (http://genetics.bwh.harvard.edu/pph/) database, all three C-terminal mutations were predicted to be damaging to the function and structure of the protein (Table 2).

Association of acquired CEBPAdm 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 CEBPAdm 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).

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

Finally, 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.

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

Gene expression profiling (GEP) was performed in a subset of the CN-AMLs patients and also includes cytogenetically abnormal patients (Table S1; n=674). Unsupervised analyses, i.e. by computing pair-wise Pearson's correlation coefficients of 674 AML cases, 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 [18]. 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).

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

The previously predictive *CEBPA*^{dm} signature [35] was hampered by the recently reported *CEBPA* silenced AML cases that carry a similar GEP [34]. 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) [213,214], which selects discriminative probesets between the classes, *CEBPA*^{dm} (n=26) and all other AML cases, *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 material 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 did not have a consistent gene expression pattern and were different from the *CEBPA*^{dm} group.

DISCUSSION

Here, we established the value of *CEBPA*^{dm} mutation in a large cohort of CN-AML patients from AMLSG and HOVON-SAKK treatment trials. Applying dHPLC and whole-gene sequencing, 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 [35,45,46,48].

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 [47].

Compared to previous studies [35,45-48], 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.

We have previously derived gene expression signatures that predict AML with inv(16), t(15;17) and t(8;21) with 100% accuracy. Here, we generated a refined GEP signature, consisting out of 25-probesets that predict *CEBPA*^{dm} AML cases (six genes overlapped with the previous signature [35], indicated in supplementary). This signature showed sensitivity and specificity of 100% and has a better predictive power than the *CEBPA*^{dm} signature that we defined before [35]. 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* [34]. Classification results were not affected by homozygous N- or C-terminal *CEBPA*^{dm} mutations or those due to germline mutation. Since this 25-probeset signature was optimized for classification it does not necessarily provide insight into the biological meaning of *CEBPA*^{dm} mutations.

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. Four of these cases developed *CEBPA*^{dm} AML, i.e., these cases acquired a mutation in the second allele. This finding is in line with previous data [215,216]. Interestingly, we for the first time identified three C-terminal germline mutations. Two of these C-terminal mutated germline 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 three 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) classification AML, "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. Conflict-of-interest disclosure: B.L., R.D., and P.J.M.V. have declared ownership interests in Skyline, a spin-off company of Erasmus University Medical Center (Erasmus MC), held in a Special Purpose Foundation of Erasmus MC. The remaining authors declare no competing financial interests.

CONTRIBUTION

E.T. performed research, data analysis, data interpretation, creation of the figures 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.

FIGURE LEGENDS

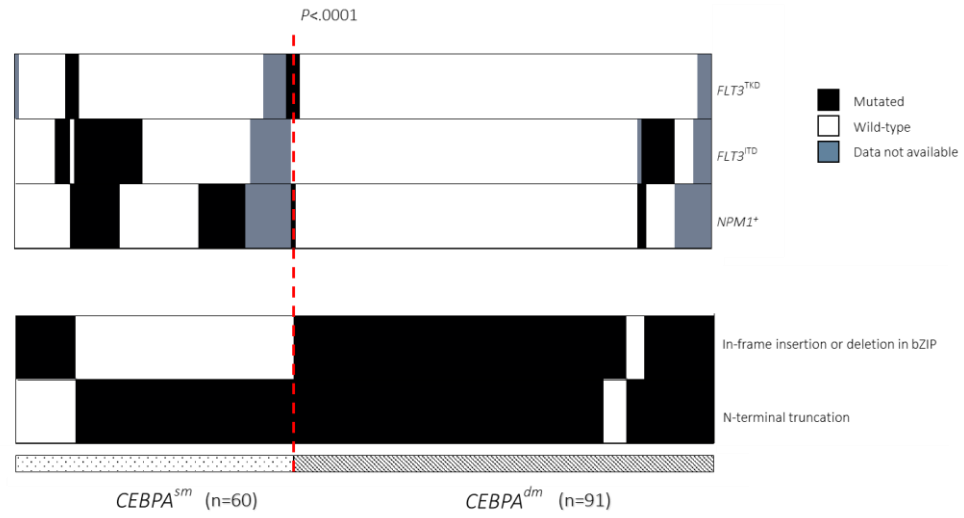


Figure 1. Distribution of concurrent mutations in *CEBPA*^{dm} and *CEBPA*^{sm} patients. Columns represent patients and rows the genotypes *FLT3*^{TKD}, *FLT3*^{ITD} and *NPM1*^{mutant} (black), wild-type (white) or missing (grey). The in-frame insertion or deletion in bZIP and N-terminal truncation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

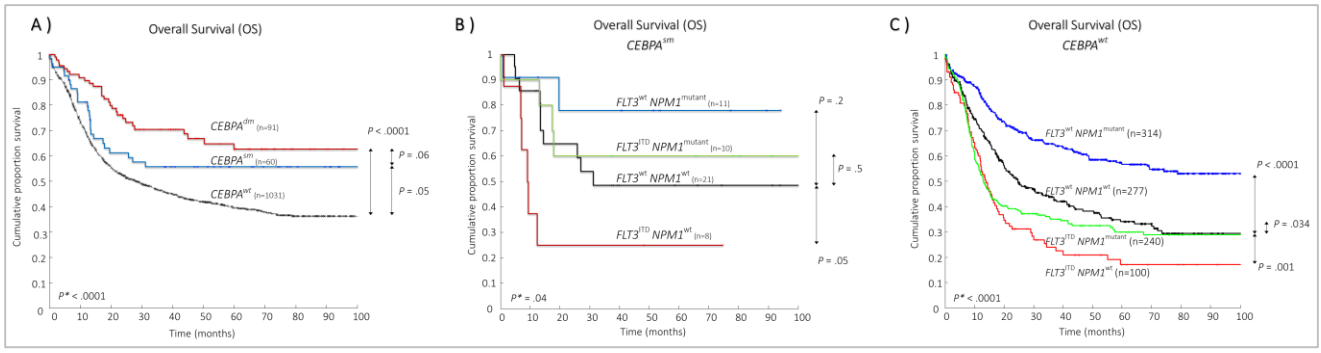


Figure 2. Kaplan-Meier survival curves of overall survival. (A) Kaplan-Meier survival curves for overall survival (OS) among the three groups *CEBPA*^{dm}, *CEBPA*^{sm} and *CEBPA*^{wt}. (B) Kaplan-Meier survival curves for OS of the four genotypes *FLT3*^{ITD}/*NPM1*^{mutant}, *FLT3*^{ITD}/*NPM1*^{wt},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

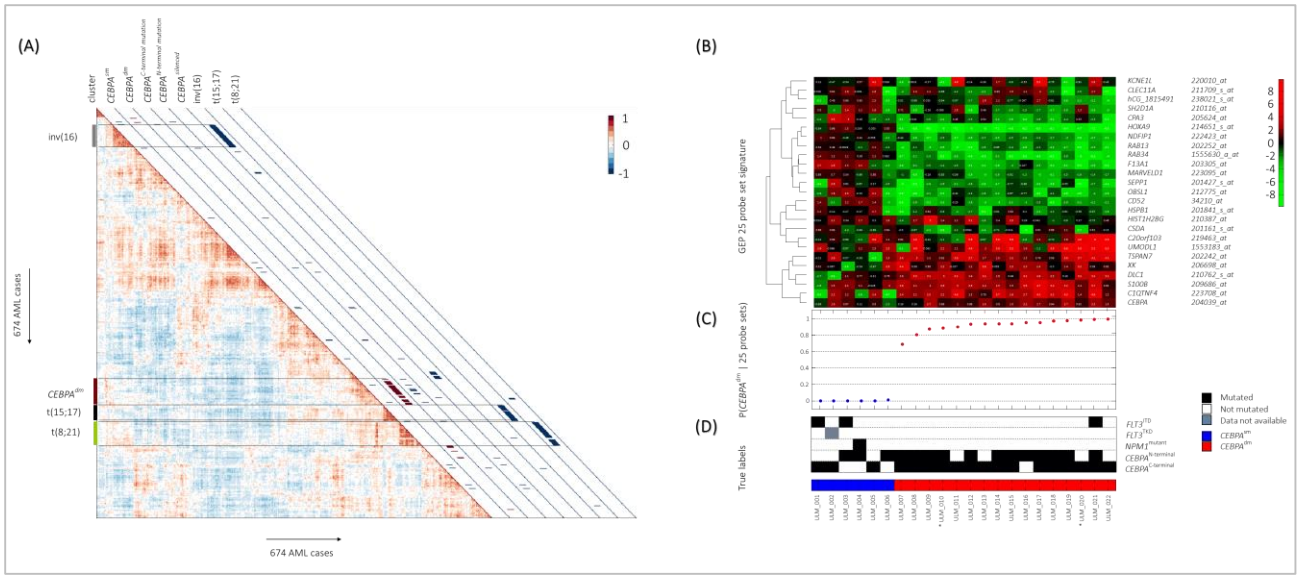


Figure 3. Unsupervised analyses and classification results of candidate *CEBPA*^{dm} cases with their gene expression profile and their molecular characteristics. (A) Pair-wise correlations between the 674 AML cases (Table S1). The cells in the visualization are colored by Pearson's correlations values, depicting higher positive (red) or negative (blue) correlations, as indicated by the scale bar. *CEBPA*^{sm}, *CEBPA*^{dm}, *CEBPA*$^{C\text{-terminal mutation}}$, *CEBPA*$^{N\text{-terminal mutation}}$, *CEBPA*^{silenced}, together with inv(16), t(15;17) and t(8;21) cases are depicted on the diagonal with a red or blue colored bar. *CEBPA*$^{C\text{-terminal mutation}}$ and *CEBPA*$^{N\text{-terminal mutation}}$ indicates the presence of homozygous mutations. (B) Candidate *CEBPA*^{dm} patients and the unambiguous *CEBPA*^{sm} patients. The expression levels are defined by the 25-probeset signature. The colors of the hierarchical clustering are relative to the mean. (C) Computed posterior probabilities, indicating the prediction of a *CEBPA*^{dm} case, given the 25 predictive probeset signature: *P*(*CEBPA*^{dm} | 25-probesets). The ordering of patients is based on the classification probabilities. The true labels (molecular characteristics) are depicted in (D). Black indicates the mutation status in *CEBPA* (*CEBPA*^{dm} or *CEBPA*^{sm}), *NPM1* (*NPM1*^{mutant}), or *FLT3* (TKD or ITD), white represents no mutation in the particular patient and a missing value is depicted in grey. The asterisk indicates the germline *CEBPA*^{dm} cases.

Characteristic	*CEBPA*^{wt} (n = 1031)	*CEBPA*^{sm} (n = 60)	*P*, *CEBPA*^{sm} vs *CEBPA* wt	*CEBPA*^{dm} (n = 91)	*P*, *CEBPA*^{dm} vs *CEBPA* wt	*P*, *CEBPA*^{sm} 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		

Table 1. Patient demographics and clinical and molecular characteristics of *CEBPA*^{wt}, *CEBPA*^{sm}, and *CEBPA*^{dm} CN-AML cases. 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).

Patient ID	Age at diagnosis, y	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_1123ins1075_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}

Table 2. Germline patient demographics and molecular characteristics. Characteristics of 5 of 71 (7%) *CEBPA-mutant* AML patients who carried *CEBPA* germline mutations. CBL indicates Casitas B-lineage lymphoma; *KRAS*, Kirsten Rat sarcoma; *NRAS*, neuroblastoma Rat sarcoma; RUNX1, runt-related transcription factor 1; and WT1, Wilms tumor 1. *Data according to GenBank accession no. Y11525. †Patients 98A-751, 07/04-268 (ULM_10), 98A-543, and 07/04-48 (ULM_20) were screened for *FLT3*^{ITD}, *FLT3*^{TKD}, *NPM1*, *NRAS*, *KRAS*, *WT1*, *RUNX1*, and *CBL* mutations; patient BioID 769 was analyzed for *FLT3*^{ITD}, *FLT3*^{TKD}, and *NPM1* mutations.

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*

Table 3. Multivariate analysis for overall survival (OS), event-free survival (EFS) and relapse-free survival (RFS) in CN-AML. Stratified Cox's proportional hazard model for multivariable analyses of *CEBPA*^{dm} and *CEBPA*^{sm} as prognostic marker for overall survival, event-free survival and relapse-free survival. Analyses included 1182 cytogenetically normal acute myeloid leukemia (CN-AML) patients with age ≤ 60. Abbreviations: HR, hazard ratio; CI, confidence interval; *FLT3*^{ITD}, *FLT3* Internal Tandem Duplications; *FLT3*^{TKD}, *FLT3* Tyrosine Kinase Domain.

**P-value* ≤ 0.05
$^{\alpha}$ *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

SUPPORTING MATERIAL

CEBPA mutation screening

Blast cells were purified using Ficoll-Hypaque (Nygaard, Oslo, Norway) centrifugation. Patients from the AMLSG-cohort (n=989) has been identified by direct sequencing, allowing the detection of homozygous mutations whereas the The 193 patients from HOVON-SAKK cohort have been pre-screened by dHPLC as previously described [35]. However, the C amplicon covering the region encoding the C-terminus of *CEBPA* was split into two smaller amplicons using two additional primers (C1B rev 5'-ACTTCTTGGCCTTGCCCGCG-3' and C2 fw 5'-CCTCCGCGCGAGTGGCGGCA-3'). Amplicon C1 was generated using primer set C fw and C1B rev and amplicon C2 with primer set C2 fw and C rev. This dHPLC strategy using these 5 *CEBPA* amplicons (fragment, A, B, C, C1 and C2) has been validated on a cohort of 550 AML cases. All known insertion/deletion and point mutants were detected. Mutations in *CEBPA* in AML patients from the AMLSG-cohort (N=989) have been identified by sequencing and confirmed by the above mentioned dHPLC strategy.

Creation and evaluation of the *CEBPA*^{dm} predictive signature

For the classification procedures we used a logistic regression model with Lasso regularization [2, 3], which selects discriminative probesets between the classes, *CEBPA*^{dm} (n=26) and 494 remaining cases (*CEBPA*^{wt} and *CEBPA*^{sm}) (HOVON-SAKK cohort). To determine the optimal signature (25-probesets, Figure 3B) we proceed with a 10-fold cross-validation where we optimized the cross-validated likelihood. With this signature we were able to discriminate the *CEBPA*^{dm} cases from *CEBPA*^{sm}, *CEBPA*^{silenced} [4], and *CEBPA*^{wt} cases. This 25-probeset signature showed overlap with six genes (*DLC1*, *MARVELD1*, *NDFIP1*, *RAB13*, *RAB34*, *UMODL1*) with the previous signature [1]. More details about the derived 25-probesets can be found in Supplementary material Table S2.

Furthermore, we inferred an optimal signature from the combined datasets (HOVON-SAKK and AMLSG-cohort). By increasing the number of training samples we introduce more information and therefore better estimations of the variance into the model. A downside is that no estimation of the test error could be inferred. For this reason we subject ourselves to the use of estimated statistics. First, we make use of the globaltest [5] which has the following null hypothesis:

H0: There is no information/pattern in the given data related to the outcomes (i.e. class labels).

When applied to the HOVON-SAKK and AMLSG-cohort, the null hypothesis can clearly be rejected (P= $2.6x10^{-18}$, P= $7x10^{-7}$ respectively). Finally, the globaltest is applied to the combined dataset and showed that the addition of the AMLSG-cohort introduced valuable information with respect to the classification of *CEBPA*^{dm} (P= $2.1x10^{-32}$). Using 10-fold cross-validation, we determined an optimal signature containing 36 probesets (Supplementary material Table S4) which attained a slightly lower estimated test error (.014 instead of .018), based on the cross-validation of the training sample, when compared to the 25-probeset signature.

Figure Legends

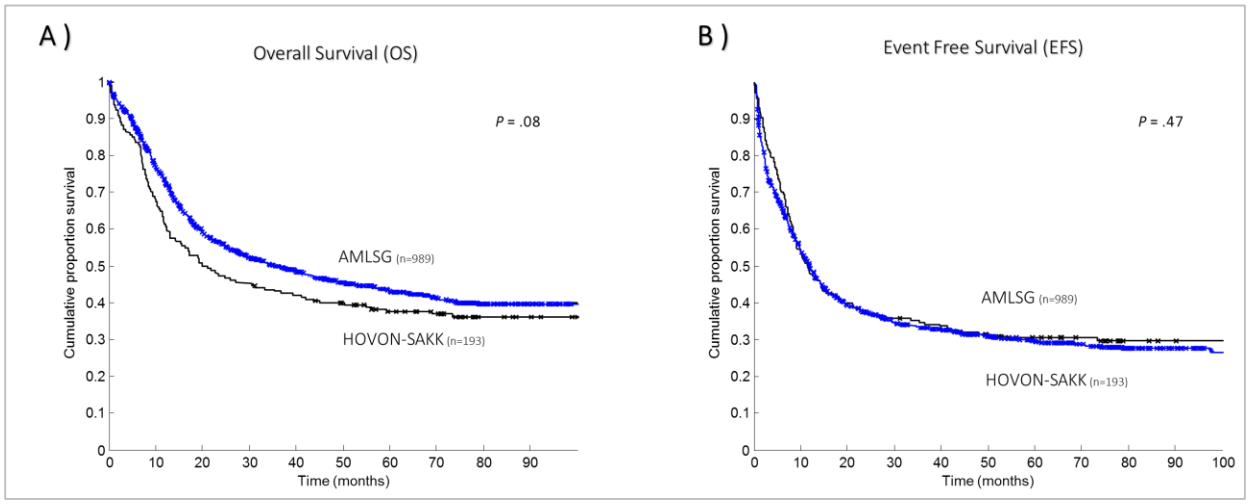


Figure S1. Kaplan-Meier survival curves of HOVON-SAKK and AMLSG-cohort. Kaplan-Meier survival curves for OS (A) and EFS (B) based on 193 and 989 patients for HOVON- SAKK- and AMLSG-cohort respectively. Log-rank test was assessed to indicate the significance between the two cohorts.

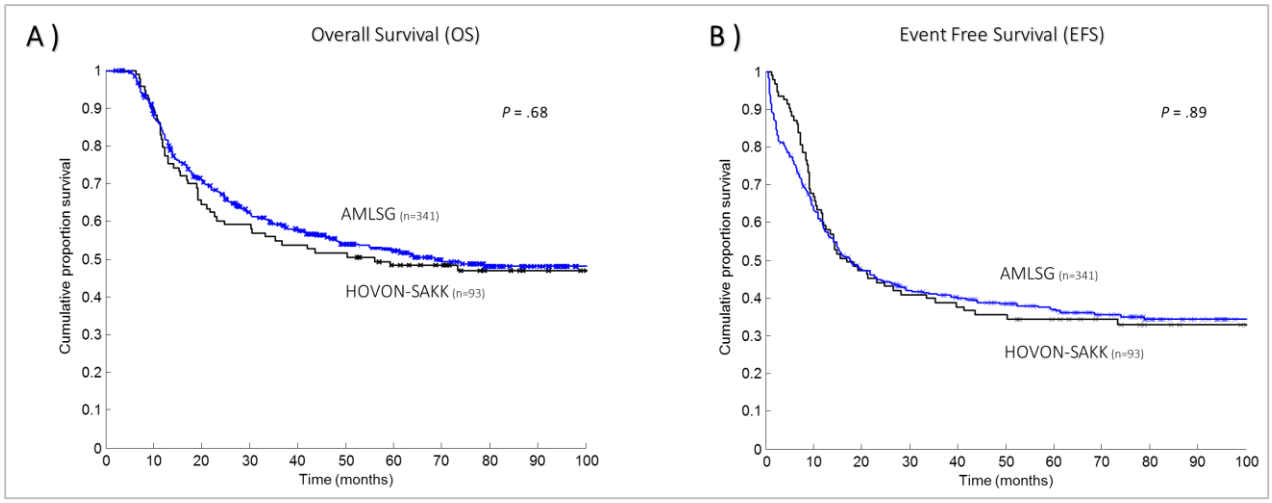


Figure S2. Kaplan-Meier survival curves of HOVON-SAKK and AMLSG-cohort with respect to therapy. Kaplan-Meier survival curves for OS (A) and EFS (B) based on 93 and 341 patients who were treated with autologous or allogeneic hematopoietic stem cell transplantation in HOVON-SAKK- and AMLSG-cohort respectively. Log-rank test was assessed to indicate the significance between the two cohorts with respect to therapy.

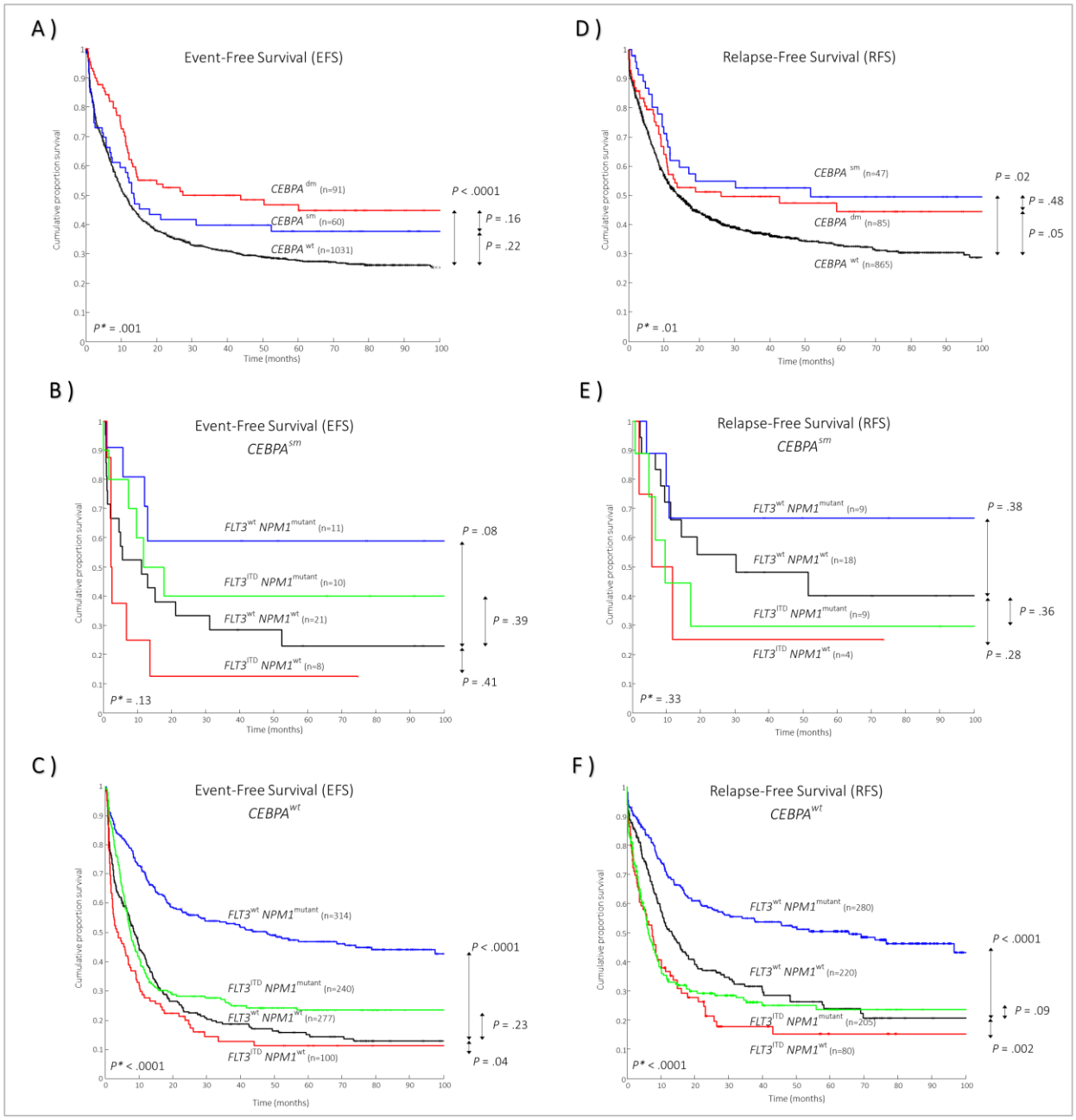


Figure S3. Kaplan-Meier survival curves of EFS and RFS. (A) and (D): Kaplan-Meier survival curves of EFS and RFS are shown for the three groups: *CEBPA*^{dm}, *CEBPA*^{sm} and *CEBPA*^{wt}. Stratified log-rank test was assessed to indicate the significance between the different groups. (B) and (E): Kaplan-Meier survival curves of EFS and RFS for the *CEBPA*^{sm} group by creating four subgroups: *FLT3*^{ITD}/*NPM1*^{mutant}, *FLT3*^{ITD} / *NPM1*^{wt}, *FLT3*^{wt}/ *NPM1*^{mutant}, and *FLT3*^{wt}/ *NPM1*^{wt}. Log-rank test was assessed to indicate the significance between the *FLT3*^{wt}/ *NPM1*^{wt} groups. (C) and (F): Kaplan-Meier survival curves of EFS and RFS within the *CEBPA*^{wt} group after creating the same subgroups. Log-rank test was assessed to indicate the significance between the composite genotypic *FLT3*/*NPM1* groups. The asterisk indicates the *P-value* for the global log-rank test.

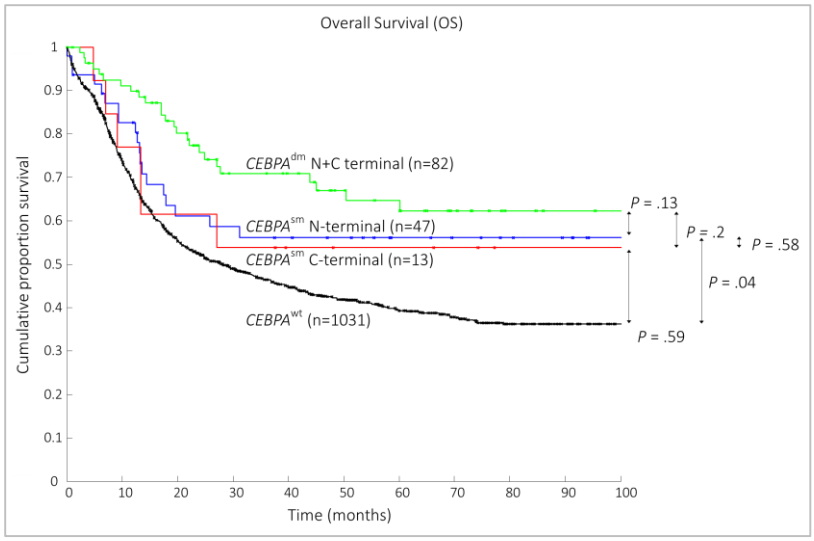


Figure S4. Kaplan-Meier survival curves of N-terminal, C-terminal *CEBPA^sm^* and N+C *CEBPA^dm^*. Kaplan-Meier survival curves for OS among the groups: N-terminal *CEBPA^sm^* (n=47), C-terminal *CEBPA^sm^* (n=13), N+C terminal *CEBPA^dm^* (n=82) and *CEBPA^wt^* (n=1031). Nine cases (homozygous *CEBPA^dm^*) are excluded. Log-rank test was assessed to indicate the significance between the different groups.

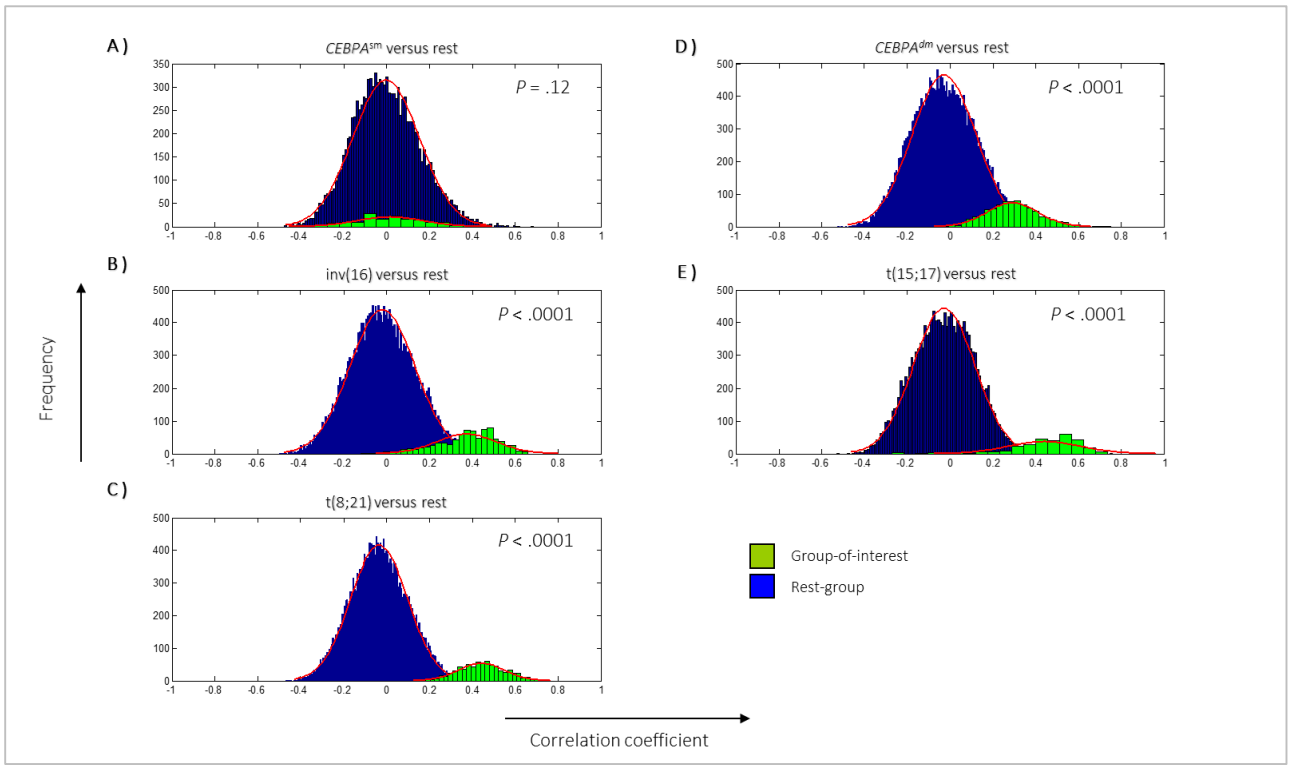


Figure S5 - Assessing significance of the detected clusters. With the use of unsupervised clustering (using Pearson's correlation coefficient) on 776 probesets, we observed that the groups; inv(16), t(15;17), t(8;21) and *CEBPA^dm^* form unique clusters whereas *CEBPA^sm^* did not. To assess the significance within the GEP cluster compared to AML cases

with any of these aberrations (rest-group), we used the Mann- Whitney-U test. High correlation is observed in (B): inv(16), (C): t(8;21), (D): *CEBPA*^{dm} and (E): t(15;17) (on average: .42, .49, .35 and .49 respectively) and differed significantly between the AML cases with any of these aberrations (P<.0001). Low correlation within the cluster is observed for *CEBPA*^{sm} (A) (on average: .15) and did not differ significantly from cases without *CEBPA*^{sm} (P=.12).

	HOVON-SAKK cohort	AMLSG-cohort	Total
Molecular data	193	989	1182
CEBPA^{dm}	18 (9.3%)	73 (7.4%)	91 (7.7%)
CEBPA^{sm}	6 (3.1%)	54 (5.5%)	60 (5.1%)
GEP data	520	154	674
CEBPA^{dm}	26 (5%)	16 (10.4%)	42 (6.2%)
CEBPA^{sm}	12 (2.3%)	6 (3.9%)	18 (2.6%)

Table S1. Molecular data and Gene Expression Profiles. Data originates from the Erasmus University Medical Center (HOVON-SAKK cohort) and the University of Ulm (AMLSG-cohort). Molecular data of 1182 CN-AML patients with age ≤ 60 is determined with a mutational screening of gene *CEBPA*. Gene Expression Profiles of 674 patients is derived using Affymetrix (Santa Clara, CA, USA) HGU133Plus 2.0 GeneChips.

Probe sets	Gene names	Regr.coef. *CEBPA*^{wt}	Regr.coef. *CEBPA*^{dm}
204039_at	*CEBPA*	-0.57251	0.57251
210762_s_at	*DLC1*	-0.27537	0.27537
223095_at	*MARVELD1*	0.19285	-0.19285
1553183_at	*UMODL1*	-0.16304	0.16304
214651_s_at	*HOXA9*	0.13178	-0.13178
1555630_a_at	*RAB34*	0.13111	-0.13111
222423_at	*NDFIP1*	0.12159	-0.12159
203305_at	*F13A1*	0.11138	-0.11138
202252_at	*RAB13*	0.10715	-0.10715
209686_at	*S100B*	-0.09809	0.09809
211709_s_at	*CLEC11A*	0.08409	-0.08409
34210_at	*CD52*	0.07837	-0.07837
206698_at	*XK*	-0.07267	0.07267
205624_at	*CPA3*	0.06188	-0.06188
201161_s_at	*CSDA*	0.05761	-0.05761
219463_at	*C20orf103*	-0.04653	0.04653
201427_s_at	*SEPP1*	0.03573	-0.03573
223708_at	*C1QTNF4*	-0.03452	0.03452
238021_s_at	*hCG-1815491*	0.02297	-0.02297
220010_at	*KCNE1L*	-0.02197	0.02197
210387_at	*HIST1H2BG*	-0.0125	0.0125
210116_at	*SH2D1A*	0.01169	-0.01169
202242_at	*TSPAN7*	-0.00961	0.00961
212775_at	*OBSL1*	0.0094	-0.0094
201841_s_at	*HSPB1//MEIS3*	0.00725	-0.00725

Table S2. 25-probeset predictive signature. The 25-probeset predictive signature is created using the logistic regression model with Lasso regularization, which determined the discriminative probesets between the classes, *CEBPA*^{dm} (n=26) versus no *CEBPA*^{dm} (n=494) (HOVON-SAKK). *CEBPA*^{wt}; CEBPA wild-type, *CEBPA*^{dm}; CEBPA double mutation, Probesets;

the 25 selected probesets, Gene names; the complementary gene names for the selected probesets, Regr.coef.; Regression coefficient derived from the Lasso procedure for the *CEBPA*^{wt} and *CEBPA*^{dm} cases.

True label	Sample ID	P(*CEBPA*^{wt} |25 probe sets)	P(*CEBPA*^{dm} |25 probe sets)
CEBPA^{sm}	ULM 001	0.99995	0.00004
CEBPA^{sm}	ULM 002	0.99985	0.00014
CEBPA^{sm}	ULM 003	0.99969	0.0003
CEBPA^{sm}	ULM 004	0.99943	0.00056
CEBPA^{sm}	ULM 005	0.99939	0.0006
CEBPA^{sm}	ULM 006	0.98616	0.01383
CEBPA^{dm}	ULM 007	0.31343	0.68656
CEBPA^{dm}	ULM 008	0.19452	0.80547
CEBPA^{dm}	ULM 009	0.12677	0.87322
CEBPA^{dm}	ULM 010	0.11421	0.88578
CEBPA^{dm}	ULM 011	0.10167	0.89832
CEBPA^{dm}	ULM 012	0.07047	0.92952
CEBPA^{dm}	ULM 013	0.06327	0.93672
CEBPA^{dm}	ULM 014	0.06322	0.93677
CEBPA^{dm}	ULM 015	0.06044	0.93955
CEBPA^{dm}	ULM 016	0.05167	0.94832
CEBPA^{dm}	ULM 017	0.04757	0.95242
CEBPA^{dm}	ULM 018	0.02838	0.97161
CEBPA^{dm}	ULM 019	0.02626	0.97373
CEBPA^{dm}	ULM 020	0.01683	0.98316
CEBPA^{dm}	ULM 021	0.01048	0.98951
CEBPA^{dm}	ULM 022	0.0054	0.99459

Table S3. Classification results. Abbreviations: True label; *CEBPA*^{sm} and *CEBPA*^{dm} are determined using sequencing and denaturing high performance liquid chromatography (dHPLC), Sample ID: Sam- ple identification number, P(*CEBPA*^{wt} |25-probesets): probability that sample ID is classified as *CEBPA*^{wt} given the 25-probeset predictive signature, P(*CEBPA*^{dm} |25-probesets): probability that sample ID is classified as *CEBPA*^{dm} given the 25-probeset predictive signature. *CEBPA*^{wt}; CEBPA wild-type, *CEBPA*^{dm}; *CEBPA* double mutation, *CEBPA*^{sm}; CEBPA single mutation.

Probe sets	Gene names	Regr.coef. *CEBPA* ^{wt}	Regr.coef. *CEBPA* ^{dm}
202007_at	*NID1*	0.01636	-0.01636
202018_s_at	*LTF*	-0.00318	0.00318
202252_at	*RAB13*	0.06454	-0.06454
202382_s_at	*GNPDA1*	0.02519	-0.02519
203305_at	*F13A1*	0.12536	-0.12536
203860_at	*PCCA*	0.04591	-0.04591
204039_at	*CEBPA*	-0.08967	0.08967
206210_s_at	*CETP*	-0.0517	0.0517
209686_at	*S100B*	-0.02699	0.02699
209905_at	*HOXA9*	0.06141	-0.06141
210298_x_at	*FHL1*	0.01507	-0.01507
210762_s_at	*DLC1*	-0.00626	0.00626
211209_x_at	*SH2D1A*	0.05816	-0.05816
211341_at	*LOC100131317 // POU4F1*	-0.00442	0.00442
211560_s_at	*ALAS2*	-0.02554	0.02554
211682_x_at	*UGT2B28*	-0.07277	0.07277
211709_s_at	*CLEC11A*	0.06346	-0.06346
212062_at	*ATP9A*	0.00432	-0.00432
214146_s_at	*PPBP*	0.07145	-0.07145
214651_s_at	*HOXA9*	0.1766	-0.1766
214835_s_at	*SUCLG2*	0.10251	-0.10251
215772_x_at	*SUCLG2*	0.0422	-0.0422
217800_s_at	*NDFIP1*	0.00914	-0.00914
219463_at	*C20orf103*	-0.09278	0.09278
220807_at	*HBQ1*	-0.03991	0.03991
222288_at	*PPP4R2*	0.0637	-0.0637
222463_s_at	*BACE1*	0.00838	-0.00838
223708_at	*C1QTNF4*	-0.10196	0.10196
224710_at	*RAB34*	0.34355	-0.34355
229638_at	*IRX3*	0.11367	-0.11367
235099_at	*CMTM8*	0.01472	-0.01472
235289_at	*EIF5A2*	-0.00758	0.00758
235438_at	*CYP7B1*	-0.23982	0.23982
235818_at	*VSTM1*	-0.05598	0.05598
239791_at	*LOC100130740*	0.08214	-0.08214
34210_at	*CD52*	0.13627	-0.13627

Table S4. 36 probeset predictive signature. The 36 probeset predictive signature is created using the logistic regression model with Lasso regularization, which determined the discriminative probesets between the classes, *CEBPA*^{dm} (n=42) versus no *CEBPA*^{dm} (n=632) (HOVON-SAKK and AMLSG-cohort). *CEBPA*^{wt}; *CEBPA* wild-type, *CEBPA*^{dm}; *CEBPA* double mutation, Probesets; the 36 selected probesets, Gene names; the complementary gene names for the selected probesets, Regr.coef.; Regression coefficient derived from the Lasso procedure for the *CEBPA*^{wt} and *CEBPA*^{dm} cases.

CHAPTER 8

The Value of Allogeneic and Autologous Hematopoietic Stem Cell Transplantation in Prognostically Favorable Acute Myeloid Leukemia with Double-Mutant *CEBPA*

BLOOD

Juli 2013 | Volume 122 | Issue 9 | doi: 10.1182/blood-2013-05-503847

The Value of Allogeneic and Autologous Hematopoietic Stem Cell Transplantation in Prognostically Favorable Acute Myeloid Leukemia with Double-mutant *CEBPA*

Erdogan Taskesen*, Richard F. Schlenk*, Yvette van Norden, Jürgen Krauter, Arnold Ganser, Lars Bullinger, Verena I. Gaidzik, Peter Paschka, Andrea Corbacioglu, Gudrun Göhring, Andrea Kündgen, Gerhard Held, Katharina Götze, Edo Vellenga, Juergen Kuball, Urs Schanz, Jakob Passweg, Thomas Pabst, Johan Maertens, Gert J. Ossenkoppele, Ruud Delwel, Hartmut Döhner H, Jan J. Cornelissen, Konstanze Döhner* and Bob Löwenberg*

*These authors corresponded equally

KEY POINTS:

- In AML with bialleleic *CEBPA*-mutant relapse-free survival was improved by allogeneic and autologous hematopoietic stem cell transplantation.
- In relapsed patients second complete remission rate was high and survival was favorable after an allogeneic transplantation.

ABSTRACT

The clinical value of allogeneic and autologous hematopoietic stem cell transplantation (alloHSCT, autoHSCT) in the subtype of acute myeloid leukemia (AML) with double-mutant *CEBPA* (*CEBPA*dm) has remained unsettled.

Among 2983 patients analyzed for *CEBPA* mutational status (age 18-60 years) treated on four HOVON/SAKK and three AMLSG protocols; 124 had AML with *CEBPA*dm and achieved first complete remission (CR1). Evaluation of the clinical impact of alloHSCT and autoHSCT versus chemotherapy was performed by addressing time dependency in the statistical analyses.

Thirty-two patients proceeded to alloHSCT from a matched related (MRD, n=29) or matched unrelated donor (MUD, n=3) and 20 to autoHSCT in CR1; 72 received chemotherapy. Relapse-free survival (RFS) was significantly superior in patients receiving an alloHSCT or autoHSCT in CR1 as compared to chemotherapy ($p<0.001$), whereas overall survival (OS) was not different ($p=0.12$). Forty-five patients relapsed. Of 42 patients treated with reinduction therapy, 35 achieved a second CR (83%) and most (n=33) patients received an alloHSCT (MRD, n=11; MUD, n=19; haplo-identical donor, n=3). Survival of relapsed patients measured from date of relapse was 46% after 3 years.

Adult AML patients with *CEBPA*dm benefit from alloHSCT and autoHSCT; relapsed patients still have a favorable outcome after reinduction followed by alloHSCT.

INTRODUCTION

Acute myeloid leukemia with mutated CCAAT/enhancer binding protein alpha (AML with mutated *CEBPA*) gene represents a provisional disease entity in the current World Health Organization classification in the category "AML with recurrent genetic abnormalities".[1,2] However, multiple studies demonstrated that AML with double-mutant *CEBPA* (*CEBPA*dm) could be clearly distinguished from AML with single mutant *CEBPA* with respect to biological and prognostic features.[3-9] In the majority of AML with *CEBPA*dm, one allele is affected by an N-terminal mutation and the second allele carries the mutation in the C-terminus (bZIP), whereas in AML with single mutant *CEBPA*, mutations occur either in the N-terminus or in the C-terminus of the gene. The previously shown favorable impact of mutant *CEBPA* in various independent comprehensive studies on prognosis[9] has more recently been specifically related to the subtype of AML with *CEBPA*dm.[3-8]

The incidence of AML with mutated (single and double) *CEBPA* ranges from 7.5% to 11% of all AML patients and from about 13% to 18% in AML exhibiting a normal karyotype.[3-8,10-13] Furthermore, the incidence of AML with mutated *CEBPA* in patients above the age of 60 years range from 8.5%[14] to 18%.[15]

Young and middle aged adults (age 18-60 years) with AML and mutated *CEBPA* and especially those with *CEBPA*dm have a comparatively high probability of achieving a complete remission after standard "7+3" induction therapy with remission rates exceeding 90%.[6,8] Treatment outcome data revealed a favorable prognosis with an overall survival after 5-years ranging between 50% and 70%[3-8] including different types of consolidation therapy with intensive chemotherapy, autologous and allogeneic hematopoietic stem cell transplantation (autoHSCT and alloHSCT). However, relapse still remains the major cause of treatment failure occurring mainly within the first 2 years after achieving a complete remission (CR). This has for instance raised the question whether autoHSCT and alloHSCT in first CR should be recommended in patients with this genetic abnormality. So far, analyses according to the type of postremission treatment in *CEBPA*dm AML patients have not become available mainly due to limited patient numbers precluding informative statistical analyses. Thus, it remains still unclear whether the favorable prognosis of AML with *CEBPA*dm can be attributed to the mutation itself irrespective of the type of applied postremission therapy (i.e. prognostic marker) or whether the favorable prognosis is the result of a high rate of cure after autoHSCT and alloHSCT in first CR and after relapse (i.e. predictive marker). Informative insight into these factors could be of direct clinical relevance as it may guide treatment decisions on the application of autoHSCT and alloHSCT already in first CR or alternatively to hold back on these approaches and reserve the option especially of an alloHSCT as salvage only in relapsed patients. To address this question, the Dutch-Belgian-Swiss HOVON/SAKK and the German-Austrian AMLSG leukemia cooperative groups performed in a joint effort an individual-patient based meta-analysis focusing on the AML *CEBPA*dm subtype in first CR. The aim was to evaluate different postremission strategies with major focus on the comparison between alloHSCT, autoHSCT and intensive chemotherapy in a large series of *CEBPA*dm AML patients in first CR. Furthermore, in an integrated approach we also included treatment after relapse and its impact on outcome.

PATIENTS AND METHODS

Patients and Treatment

All patients included in this study were recruited within two major leukemia cohorts. AML patients from Cohort I (n=3450) were enrolled in the Dutch-Belgian-Swiss Hemato-Oncology Cooperative Group (HOVON/SAKK) trials HOVON04(A), HOVON29/SAKK30/95[17-19], HOVON42(A)/SAKK30/00 and HOVON92/SAKK30/08[18-20] (www.hovon.nl). Patients received two successive cycles of anthracycline-cytarabine and amsacrine-cytarabine based remission induction chemotherapy and, subsequently, in first CR consolidation chemotherapy, autoHSCT after myeloablative therapy according to a randomization against chemotherapy and depending on an adequate stem cell collection[20], or alloHSCT following mainly myeloablative conditioning depending on the availability of a matched related donor.

Cohort II (n=2274) comprised patients who were enrolled in the German-Austrian AML Study Group (AMLSG) trials AML HD93,[21] AML HD98A[22] and AMLSG 07-04 (ClinicalTrials.gov Identifier: NCT00151242). Consistently throughout all AMLSG trials, patients with AML exhibiting an intermediate-risk karyotype with mutant *CEBPA* were intended to receive, i*)* a double induction therapy with ICE (idarubicin, cytarabine, etoposide) and, ii*)* repetitive cycles of high-dose cytarabine based consolidation therapy or, if an HLA-matched family donor was available, an allogeneic HSCT after a myeloablative conditioning regimen.

Patients were selected from the total cohort, if they fulfilled all three of the following criteria, i) normal karyotype or intermediate-risk karyotype according to ELN criteria[2], ii) *CEBPA*dm, iii) CR after induction therapy. The selection process is illustrated in Figure 1. In 90% (5147/5724) of the patients information on cytogenetics was available. In these 5147 patients the *CEBPA* mutation status was available in 2983 (58%); of those 137 exhibited a *CEBPA*dm in the context of a normal karyotype or intermediate-risk cytogenetics (4.6%).One-hundred-twenty-four achieved a first CR after induction therapy within the different protocols (90.5% CR rate) and were included into this study.

All patients provided written informed consent in accordance with the Declaration of Helsinki. All trials were approved by the Institutional Review Boards. *CEBPA* mutational status was identified by denaturing High-Performance Liquid Chromatography (dHPLC), PCR amplification followed by direct sequencing or fragment-length analysis (GeneScan) and subsequent sequence analysis in any positive cases.[4,5] Cytogenetics and molecular analyses were performed as described before.[8,10,23-25]

Statistical analysis

The definition of CR and survival endpoints such as overall survival (OS), cumulative incidence of relapse (CIR) and death (CID), as well as relapse-free survival (RFS) were based on the recommended consensus criteria.[2] Actuarial estimates were used for assessment of the median follow-up for survival. Patient characteristics were compared by the Kruskal-Wallis test (continuous variables) and the Fisher's exact test (categorical variables). Cumulative incidence of relapse (CIR) and death (CID) were analyzed according to the method of Gray.[26] To address the time dependence

of the variables alloHSCT and autoHSCT, the graphical representation according to the method of Simon and Makuch was used as well as the Mantel-Byar test, as appropriated statistical approach in univariable analyses.[27,28] For multivariable analyses an extended Cox regression model was used according to the method of Andersen and Gil.[29] For all analyses, a *P-value* was considered statistically significant if it was less or equal than .05. All statistical analyses were performed using the statistical software Stata Statistical Software, Release 12.

RESULTS

Demographics and clinical baseline characteristics of the study population

In this cooperative individual-patient data meta-analysis, 124 *CEBPA*dm AML patients were included with normal karyotype or intermediate-risk cytogenetics, age between 18 and 60 years, and first CR after induction therapy. The patients were selected from the total study population treated in HOVON/SAKK and AMLSG prospective multicenter clinical trials recruited between 1987 and 2009 (Figure 1). Baseline characteristics and demographics for the total cohort are shown in Table 1.

No significant difference in overall survival was seen between HOVON/SAKK and AMLSG *CEBPA*dm patient cohorts (n=50 vs. n=74, Cox test, *P*=.36); molecular and clinical variables were comparable between HOVON/SAKK and AMLSG *CEBPA*dm patient cohorts with exception of platelet counts (*P*=.02, lower in AMLSG) and bone marrow blasts (*P*<.0001, lower in HOVON/SAKK).

Postremission therapy, Cumulative Incidence of Relapse (CIR) and Death in CR (CID) in *CEBPA*dm patients

Distribution of postremission treatment modalities in the 124 patients was as follows: alloHSCT, n=32 (matched related donor [MRD] n=29, matched unrelated donor [MUD] n=3); autoHSCT, n=20; intensive chemotherapy, n=72. The median time interval from diagnosis to achievement of first CR was 1.1 months (range 0.36 - 4.11) with a trend (p=0.053) for a longer interval in patients who received an autoHSCT as postremission treatment (median, 1.25 months) compared to those who received an alloHSCT (median, 1.1 months) or chemotherapy (median, 1.1 months). The median time interval from first CR to alloHSCT and autoHSCT was 4.0 months (range 0.8-6.5) and 2.3 months (range 0.4-6.2), respectively.

In total, 45 relapses (one after alloHSCT, 5 after autoHSCT, and 39 after intensive chemotherapy) and 19 treatment-related deaths (7 after alloHSCT, 3 after autoHSCT, and 9 after intensive chemotherapy) after a median time measured from CR1 of 10.8 months (range, 4.4-61) and 15 months (range, 1.2-144) occurred. This leads to estimates of CIR and CID after 5-years of 3% (SE 3%) and 24% (SE 8%) for alloHSCT, 27% (SE 11%) and 13% (SE 9%) for autoHSCT, and 58% (SE 6%) and 10% (SE 4%) for intensive chemotherapy, respectively.

Treatment after relapse

After relapse, 41 patients received intensive reinduction therapy, one patient had repetitive cycles of subcutaneous azacitidine and three patients were treated with supportive care only. The second CR rate in all relapsed patients was 78% (35/45) and in those receiving reinduction therapy (including azacitidine) 83% (35/42). Thirty-three patients received an alloHSCT (MRD n=11, MUD n=19, haplo-identical donor n=3) after reinduction therapy. At the time of alloHSCT, 28 patients achieved a second CR after reinduction therapy, four had refractory disease and one patient who relapsed after alloHSCT in first CR, received a stem cell boost from the same donor in second CR. Only one patient was treated with an autoHSCT after relapse.

Survival analysis

The median follow-up time of patients still alive at the date of last contact was 62 months. RFS and OS of the whole *CEBPA*dm patient cohort after 5-years were 48% (95%-CI, 38-57%) and 63% (95%-CI, 53-72%), respectively. There was no difference in RFS (p=0.24) and OS (p=0.87) between patients exhibiting a normal karyotype and those with intermediate-risk karyotypes (Table 2, Figure 2). Based on the Mantel-Byar test including alloHSCT and autoHSCT applied in first CR as time dependent variables, RFS is significantly superior in patients receiving an HSCT (p<0.001), with significant differences in favor of alloHSCT and autoHSCT as compared to intensive chemotherapy (p<0.001 and p=0.019, respectively) (Figure 3a). Multivariable analysis based on an Andersen-Gill model including time dependent postremission strategy as well as pretreatment values (Table 1) of WBC, platelets, BM-blast percentage, age and karyotype (normal versus abnormal) revealed that alloHSCT (HR, 0.23; p<0.001) and autoHSCT (HR, 0.37; p=0.012) applied in first CR had independently a favorable prognostic impact with regard to RFS (Table 3). However, apparently due to a high second CR rate after salvage therapy, the superior RFS after alloHSCT and autoHSCT did not translate into a better OS in univariable (p=0.12) (Figure 3b) and multivariable analyses (Table 4).

In 45 relapsed patients OS measured from the date of relapse was 46% (95%-CI, 30-60%) after 3 years (Figure 4). All patients surviving more than 2 years after first relapsed had undergone an allogeneic HSCT (n=15, Figure 4).

DISCUSSION

This report focuses on the evaluation of the clinical impact in *CEBPA*dm patients of alloHSCT, autoHSCT in comparison to intensive consolidation therapy in first CR as well as on the impact of reinduction chemotherapy and alloHSCT after relapse. To this attempt, we report on 124 adults with AML and a normal karyotype or intermediate-risk karyotypes that harbor a *CEBPA*dm, and are aged 60 years and below. Only one relapsed patient received an autoHSCT in second CR and therefore we were not able to evaluate the clinical impact of autoHSCT after relapse. Our results clearly show that adult AML patients with the *CEBPA*dm genotype significantly benefit from alloHSCT and autoHSCT in first CR with respect to RFS.

In recent years it has become apparent that the favorable prognosis of *CEBPA* gene mutations largely depends on the presence of the *CEBPA*dm mutation type. At the molecular level AML with the *CEBPA*dm compared to AML with *CEBPA*

single mutation type is associated with a lower frequency of coexisting *NPM1* mutations and *FLT3* internal tandem duplications (ITD).[6,8] Therefore, several investigators have recently suggested to restrict the provisional WHO 2008 entity AML with *CEBPA* mutations to those with biallelic mutations.[4-8] As a direct consequence the incidence of this AML entity defined by *CEBPA*dm decreases by about 40%[5,8] to a frequency of 3% to 6% of all AML cases. The low frequency of *CEBPA*dm explains why comparative analyses with regard to different postremission strategies such as alloHSCT, autoHSCT and intensive chemotherapy have so far not been performed. Beside the recommended ELN-risk category[2] AML with CEBPA we also included into the analyses the group of intermediate-risk karyotypes which includes approximately 30% of patients with chromosomal abnormalities, in particular interstitial deletion 9q[30] and 11q (Table 2) instead of only patients with AML exhibiting a normal karyotype. This approach is supported by the similar favorable outcome in AML with *CEBPA*dm with and without normal karyotype in uni- and multivariable analyses presented here (Figure 2). Thus, these data add evidence that AML with *CEBPA*dm may be regarded as a distinctive AML entity irrespective of additional chromosomal abnormalities categorized within the cytogenetically defined intermediate-risk group.[2]

In the current analyses we started with a total cohort of 5724 patients from which in 5147 a karyotype and of those in 2983 the *CEBPA* mutational status was available. This large cohort was required to finally achieve a sufficiently high number of 124 AML patients with *CEBPA*dm in first CR representing the basis of our analyses. This approach underlines that large cooperative intergroup meta-analyses are warranted to evaluate treatment effects with acceptable statistical power in rare but clinically highly relevant patient subsets.

Our patients were treated in seven different treatment trials with in part changing therapeutic concepts over time. Thus, it is impossible to apply the rigorous statistical standards for postremission treatment allocation such as up-front or the so called genetic randomization. Instead, we applied statistical methods that have all in common that group allocation is implemented as a dynamic process over time with a transition from the no-transplant group to the alloHSCT or autoHSCT groups at the time-point of HSCT. This approach reduces the time-to-treatment bias and provides a solid statistical methodology in situations where simple Kaplan-Meier plots and log-rank tests as well as simple Cox regression models are no longer valid. However, to further reduce selection bias towards HSCT in first CR our univariable comparisons were complemented by multivariable Andersen-Gil regression models addressing again the time-to-treatment bias[29] by including important pretreatment characteristics.

By using methods that adjust for the time from CR to consolidation we were able to show a clear superior RFS ($p<0.001$) in patients who received an alloHSCT or an autoHSCT in first CR of 73% (95%-CI, 54-86%) and 60% (95%-CI, 33-79%) after 5-years, respectively. These survival rates compare favorably to the RFS of 32% (95%-CI, 21-45%) in patients receiving intensive chemotherapy only. A similarly good outcome has also been reported for other AML entities categorized into the favorable ELN-risk group such as core-binding factor AML (CBF-AML) including AML with inv(16) or t(16;16) and AML with t(8;21) after both alloHSCT and autoHSCT.[31,32] However, our results suggest that RFS after intensive chemotherapy is substantially lower for AML with *CEBPA*dm as compared to that of CBF-AML.[33,34]

Nevertheless, due to a high second CR rate in reinduced relapsed patients with more than 80% and a high proportion of patients proceeding to an alloHSCT after relapse, the high relapse rate in the chemotherapy subgroup did not translate into a significant inferior OS. In fact, the high second CR rate and favorable survival after relapse observed in the study reported here are comparable to the survival probabilities observed in AML with inv(16).[33,34] Based on these data, AML with *CEBPA*dm and AML with inv(16) or t(16;16) appear as two well-defined exceptions from the general notion that after relapse a second CR is rarely achieved.[35] Therefore, instead of applying alloHSCT as the compelling option in first CR, an alternative and not unreasonable strategy would be to postpone the alloHSCT in first CR and keep the option of alloHSCT for salvage for the restricted fraction of patients after relapse.[35,36] Indeed, our data supports both strategies, alloHSCT or autoHSCT in first CR versus intensive chemotherapy as consolidation in first CR and reinduction followed by alloHSCT in case of relapse. Of note, the good results in our study with autoHSCT are paralleled by those obtained in Core-binding factor AML[31,32] indicating a specific chemo-sensitivity of these AMLs with favorable risk according to the ELN recommendations[2] to dose escalation during consolidation therapy in first CR. Patients have to be well informed about the risks and consequences of alloHSCT and autoHSCT in first CR regarding: i) short[37] and long[38] term physical and psychological impairment, ii) infertility and a higher rate of treatment-related mortality (e.g. for alloHSCT we observed in our cohort 24% at 5-years), and iii) increased rates of transplantation-related morbidity and mortality for alloHSCT when the latter is performed after relapse.[16] Besides the survival considerations there are various other arguments that would favor the choice of an autoHSCT in first CR as compared to alloHSCT. Given the nearly identical RFS and OS rates after alloHSCT and autoHSCT the focus of outcome evaluations can be broadened to consider quality of life (QoL) aspects and late effects after transplantation as well as health economics, with a significantly better QoL and fewer late effects after autoHSCT compared to alloHSCT[39] whereas data on health economics are very system specifiec.[40]

In summary, our data provide novel clinical information that may be useful for refining the WHO 2008 classification and the ELN-risk categorization for the provisional entity AML with *CEBPA*dm in that beyond normal karyotype all intermediate-risk cytogenetics should be included. From a clinical perspective alloHSCT and autoHSCT performed in first CR were associated with comparatively excellent RFS and OS, whereas the reduced rate of RFS in patients receiving consolidation with intensive chemotherapy could be made up after relapse by a high rate of second CR followed by alloHSCT. Thus the marker *CEBPA*dm develops with respect to RFS to a predictive marker indicating superior RFS after autoHSCT and alloHSCT, but remains a prognostic marker with respect to OS. The pros and cons of alloHSCT and autoHSCT during first CR or, as an alternative option, alloHSCT after relapse have to be carefully considered and discussed with the patients with possible individual adaptation of treatment recommendations taking into account the patient's personal context.

ACKNOWLEDGMENTS

We thank Jasper Koenders and Francois Kavelaars for their contribution in the collection of the data. This research was supported by the Center for Translational Molecular Medicine and by Else Kröner-Fresenius-Stiftung grant P38/05//A49/05//F03, the Network of Competence Acute and Chronic Leukemias grant 01GI9981, and the Bundesministerium für Bildung und Forschung, Germany, grant 01KG0605 ("IPD-metaanalysis: a model-based hierarchical prognostic system for adult patients with acute myeloid leukemia [AML]").

AUTHOR CONTRIBUTION

Conception and Design: Richard F. Schlenk, Erdogan Taskesen, Bob Löwenberg, Yvette van Norden, Hartmut Döhner, Konstanze Döhner, Ruud Delwel. **Provision of study materials or patients:** Richard F. Schlenk, Erdogan Taskesen, Jürgen Krauter, Arnold Ganser, Verena Gaidzik, Peter Paschka, Gudrun Goehring, Andrea Kündgen, Katharina Götze, Peter Valk, Jan J. Cornelissen, Gert J. Ossenkoppele, Juergen Kuball, Urs Schanz, Edo Vellenga, Hartmut Döhner, Ruud Delwel, Konstanze Döhner, Bob Löwenberg. **Collection and assembly of data:** Richard F. Schlenk, Erdogan Taskesen, Yvette van Norden, Jürgen Krauter, Arnold Ganser, Verena Gaidzik, Peter Paschka, Gudrun Goehring, Andrea Kündgen, Gerhard Held, Katharina Götze, Peter Valk, Jan J. Cornelissen, Gert J. Ossenkoppele, Juergen Kuball, Urs Schanz, Edo Vellenga, Hartmut Döhner, Ruud Delwel, Konstanze Döhner, Bob Löwenberg. **Data analysis and interpretation:** Richard F. Schlenk, Erdogan Taskesen, Bob Löwenberg, Yvette van Norden, Hartmut Döhner, Konstanze Döhner. **Manuscript writing:** Richard F. Schlenk, Erdogan Taskesen, Bob Löwenberg, Yvette van Norden, Hartmut Döhner, Konstanze Döhner. **Final approval of manuscript:** Richard F. Schlenk, Erdogan Taskesen, Yvette van Norden, Jürgen Krauter, Arnold Ganser, Lars Bullinger, Verena Gaidzik, Peter Paschka, Andrea Corbagioglu, Gudrun Goehring, Andrea Kündgen, Gerhard Held, Katharina Götze, Edo Vellenga, Juergen Kuball, Urs Schanz, Jakob Passweg, Thomas Pabst, Johan Maertens, Gert J. Ossenkoppele, Ruud Delwel, Hartmut Döhner, Jan J. Cornelissen, Konstanze Döhner, Bob Löwenberg

FIGURE LEGENDS

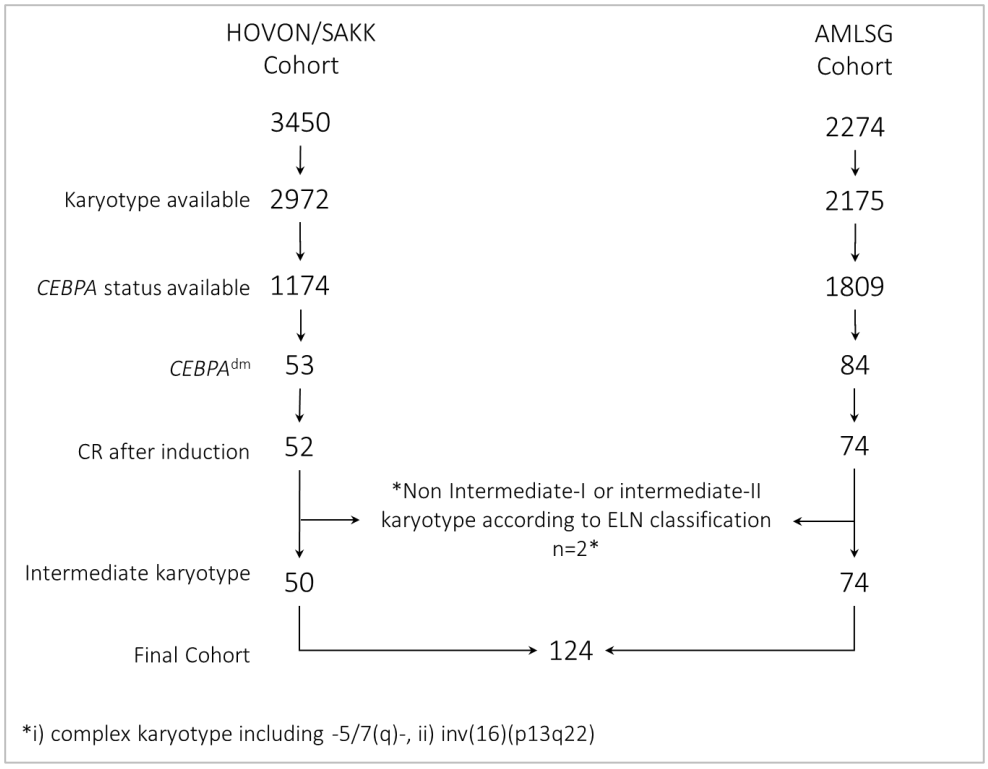


Figure 1. Flow chart on patient selection. Number of patients according to each selection step.

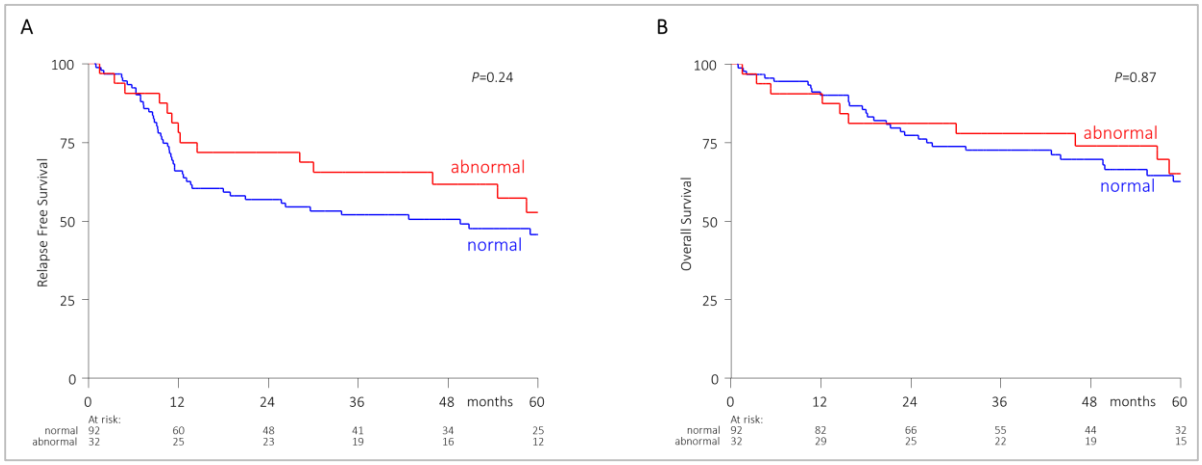


Figure 2. Influence of karyotype abnormalities on outcome. Kaplan-Meier plots for the endpoints a) RFS and b) OS according to the karyotype (normal versus abnormal).

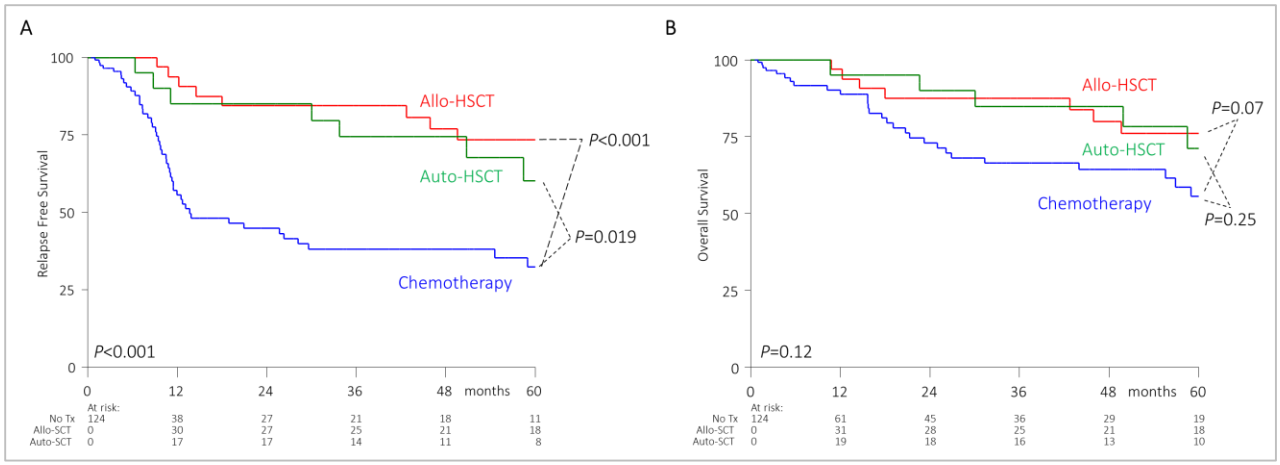


Figure 3. Influence of postremission treatment modality (alloHSCT, autoHSCT, chemotherapy) on RFS (Fig 3a) and OS (Fig 3b). Simon-Makuch plots for the endpoints a) RFS and b) OS according to type of postremission therapy; allo, allogeneic HSCT; auto, autologous HSCT; chemotherapy.

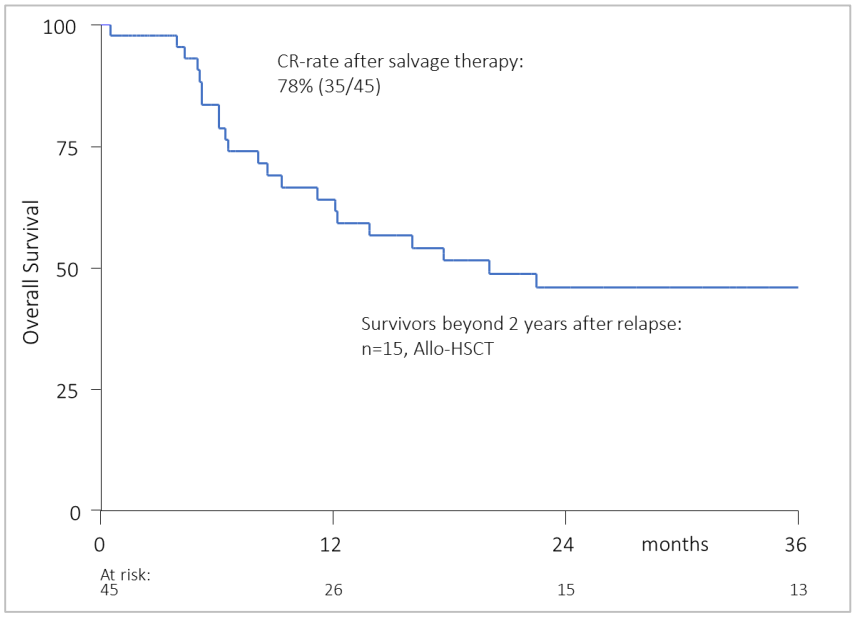


Figure 4. Outcome of patients after relapse. Simon-Makuch plot for the endpoint OS, second CR rate after salvage remission induction therapy and treatment details for patients surviving longer than 2 years.

	Total cohort	Allo-HSCT	Auto-HSCT	no-HSCT	*P*-value
Characteristics	n=124	n=32	n=20	n=72	
Age, years					
Median (range)	44 (16-60)	40	41	45	0.124
Sex, no. (%)					
Male	66 (53%)	21 (66%)	12 (60%)	33 (46%)	0.147
Female	58 (47%)	11 (34%)	8 (40%)	39 (54%)	
WBC, x 10^9/L					
Median (range)	32 (2-248)	16 (2-174)	36 (3-157)	34 (2-248)	0.228
Missing	1			1	
Platelets, x10^9/l					
Median (range)	41 (4-319)	39 (11.282)	47 (10-115)	40 (4-319)	0.421
Missing	3			3	
Bone marrow blasts, (%)					
Median (range)	75 (7-100)	71 (7-99)	77 (31-99)	75 (25-100)	0.535
Missing	5	1		4	
Type of AML, no. (%)					
De novo AML					
M0	6 (5%)	1 (3%)		5 (7%)	0.237
M1	45 (36%)	10 (31%)	9 (45%)	26 (36%)	
M2	54 (44%)	16 (50%)	6 (30%)	32 (44%)	
M4	6 (5%)	2 (6%)	2 (10%)	2 (3%)	
M5	2 (2%)	1 (3%)	1 (5%)		
M6	1 (1%)		1 (5%)		
Unclassified	1 (1%)			1 (1%)	
Missing	9 (7%)	2 (6%)	1 (5%)	6 (9%)	
Cytogenetics, no, (%)					
Normal karyotype	92 (74%)	21 (66%)	14 (70%)	57 (79%)	0.317
Molecular characteristics no., (%)					
FLT3^{ITD}	11 (9%)	5 (16%)	1 (5%)	5 (7%)	0.38
NPM1$^+$	2 (2%)	1 (3%)		1 (1%)	0.67
Missing	1		1		

Table 1. Clinical and genetic characteristics of the total cohort and according to applied postremission therapy. Abbreviations: AML, acute myeloid leukemia; WBC: white blood cell count; *FLT3*^{ITD}, *FLT3* internal tandem duplication; *NPM1*$^+$, nucleophosmin 1; Subheadings under "de novo AML" refer to French American British classification subtypes; wt, wild-type; neg, negative.

del(9q)
46,XY,del(9)(q12q31)[20]
46,XY,del(9)(q12q22),del(11)(q13q23)[21]
46,XY,del(9)(q12q34),del(11)(p11p15)[12]/46,XY[4]
46,XX,del(9)(q12q31~32)[26]/47,idem,+21[2]/46,XX[7]
46,XX,del(9)(q1?2q3?2)[7]/46,XX[14]
46,XY,del(9)(q13q22)[22]
46,XY,del(9)(q13q22)[8]/46,XY[18]
46,XY,del(9)(q13q34)[2]/46,XY[19]
46,XY,del(7)(q22q32)[30]/46,XY,del(7)(q22q32),del(9)(q13q32)[2]
46,XX,del(9)(q21q22)
46,XX,del(9)(q22)[20]
46,XY,del(9)(q22q34)[13]/46,XY[7]
46,XY,del(9)(q22q34)[10]
46,XY,del(9)(q22q34)[2]/46,XY[17]
46,XY,del(9)(q22q34)[5]/47,XY,del(9)(q22q34),+21[5]/46,XY[1]
46,XY,del(9)(q3?1)[2]/46,XY[38]
46,XX,del(9)(q3?1) or del(9)(q22q34)[11]/46,XX[19]
del(11q)
46,XX,del(11)(q13q25)[20]
46,XY,del(11)(q14q25)[3]/46,XY[60]
46,XY,del(11)(q21q23)[6]/46,XY[9]
46,XY,del(11)(q21q23)[7]
other
45,X,-Y[14]
45,X,-Y[8]/46,XY[14]
46,XX,del(1)(p32p34)[21]
46,XX,del(7)(p13p15)[4]/46,XX[6]
46,XX,iso(17)(q10)[10]/46,XX[11]
47,XX,+5[7]/46,XX[16]
47,XX,+10[18]/46,XX[2]
47,XY,+10[20]
47,XX,+21[22]
47,XY,+21[6]/46,XY[9]

Table 2. Abnormal karyotypes grouped according to leading aberrations.

Prognostic markers*	RFS		
	HR	95%-CI	*P*
Allogeneic HSCT	0.23	0.11– 0.51	<0.001
Autologous HSCT	0.37	0.17 - 0.80	0.012
Log_{10} (WBC)	1.4	0.77 – 2.56	0.27
Log_{10} (platelets)	0.81	0.40-1.66	0.57
% BM blasts (difference 10%)	0.94	0.82 – 1.08	0.4
Age (difference of 10 years)	0.86	0.68 – 1.08	0.2
Abnormal karyotype	0.73	0.38 – 1.40	0.35

Table 3. Andersen-Gill model for the endpoint relapse-free survival. Abbreviations: HSCT hematopoietic stem cell transplantation; WBC, white blood cell count; % BM blasts, percentage bone marrow blasts.

Prognostic markers*	RFS		
	HR	95%-CI	*P*
Allogeneic HSCT	0.5	0.21 – 1.17	0.11
Autologous HSCT	0.57	0.23 – 1.40	0.22
Log_{10} (WBC)	1.34	0.64 – 2.80	0.44
Log_{10} (platelets)	0.95	0.40 – 2.26	0.91
% BM blasts (difference 10%)	1.04	0.87 – 1.24	0.69
Age (difference of 10 years)	1.08	0.82 – 1.42	0.59
Abnormal karyotype	1.14	0.55 – 2.34	0.73

Table 4. Andersen-Gill model for the endpoint overall survival. Abbreviations: HSCT hematopoietic stem cell transplantation; WBC, white blood cell count; % BM blasts, percentage bone marrow blasts.

CHAPTER 9

Two Splice Factor Mutant Leukemia Subgroups Uncovered at the Boundaries of MDS and AML using Combined Gene-Expression and DNA-Methylation Profiling

BLOOD

Under review

Two Splice Factor Mutant Leukemia Subgroups Uncovered at the Boundaries of MDS and AML using Combined Gene expression and DNA-Methylation Profiling

Erdogan Taskesen, Marije Havermans, Kirsten van Lom, Mathijs Sanders, Yvette van Norden, Eric Bindels, Remco Hoogenboezem, Marcel J.T. Reinders, Maria E. Figueroa, Peter J.M. Valk, Bob Löwenberg, Ari Melnick and Ruud Delwel

KEY POINTS:

- Splice factor mutant myeloid malignancies transcend the boundaries between AML and MDS.
- Integrated analysis of as gene expression and DNA-methylation profiling in large leukemia cohort uncovers novel subtypes.

ABSTRACT

Mutations in splice factor (SF) genes occur more frequently in myelodysplastic syndromes (MDS) than in acute myeloid leukemias (AML). We sequenced cDNA from 7 human RAEB (refractory anemia with excess of blasts), 13 RAEB in transformation (RAEB-t) and 324 AML patients and determined the presence of SF-hotspot mutations in *SF3B1*, *U2AF35,* and *SRSF2*. SF-mutations were found in 2 RAEB, 5 RAEB-t and 28 AML cases. SF-mutant AMLs were older, showed lower white blood cell counts, lower marrow blast percentages and higher erythroblast percentages than SF-wild-type AMLs. Besides the blast percentages, no differences were found between SF-mutant RAEB, RAEB-t and AML cases. This suggests that these SF-mutant malignancies may be considered as myeloid malignancies that transcends the boundaries of AML and MDS. An integrated analysis of gene expression (GEP) and DNA-methylation profiling (DMP) data revealed two unique patient-clusters highly enriched for SF-mutant AML/RAEB(T). The combined GEP/DMP signatures revealed one SF-mutant subset with an erythroid signature. The other SF-mutant cluster was enriched for *NRAS/KRAS* mutations and showed an inferior survival. We conclude that SF-mutant AML/RAEB(T) constitute a related disorder overriding the artificial separation between AML and MDS, and that SF-mutant AML/RAEB(T) is composed of two molecularly and clinically distinct subgroups.

INTRODUCTION

Myelodysplastic syndromes (MDS) are characterized by a deregulation of blood cell formation and frequently develop into acute myeloid leukemia (AML). MDS patients feature recurrent somatic mutations in multiple components of the RNA splicing machinery[217-220]. These mutations are frequently seen in the splice factor (SF) genes, *SRSF2, U2AF35, ZRSR2, U2AF65, SF1, SF3B1, SF3A1 or PRPF40B*[217-219]. Among the many non-recurrent missense mutations, eight

mutational hotspots were found, i.e. in *U2AF35* (two hotspots), *SRSF2* (one hotspot) and *SF3B1* (five hotspots)[217]. Although, these SF-mutations have been reported to frequently associate with the presence of ring sideroblasts (RS)[217,221], MDS without RS can harbor SF-mutations as well [217,218]. Mutations in *SF3B1* are strongly associated with refractory anemia with ring sideroblasts (RARS), whereas in MDS without RS no association with a specific mutation was observed[217]. Some of the SF-mutations also appeared to have prognostic relevance[222-225].

The French-American-British (FAB) classification for MDS and AML has been used to delineate the transitional zone in marrow and blood blast percentages that separate MDS and AML. RAEB is defined as an MDS with ≥5% but ≤20% blasts in the bone marrow and RAEB-t as ≥5% blasts in the blood, or bone marrow blasts >20% but <30%, or the presence of Auer rods[226,227]. RAEB is considered to be an MDS subtype closely related to AML, whereas RAEB-t is classified as AML according to the WHO[228]. Thus, the distinction between RAEB, RAEB-t and AML is arbitrary and molecular abnormalities have not provided a clear basis for the separation of AML and MDS biology. It is therefore possible that subsets of RAEB, RAEB-t and AML, in particular the ones with the same class of molecular abnormalities, such as splice factor (SF)-mutations would represent one common molecular leukemia subtype. We investigated in this study the distribution of eight hotspot SF-gene mutations[217] in RAEB (N=7), RAEB-t (N=13) and AML (N=324) samples. The data revealed that splice factor (SF) mutant RAEB, RAEB-t and AML share highly similar phenotypes and suggest that these malignancies should be considered as one typical SF-mutant AML subset.

AML subtypes with unique molecular defects, such as patients with recurrent chromosomal translocations t(8;21), t(15;17), inv(16), or with mutations in *CEBPA* or in *NPM1* can be uncovered very specifically using gene expression profiling or DNA-methylation profiling (GEP[18] or DMP[58]) data, derived from large AML patient cohorts[18,24,44,121]. Application of GEP or DMP in cohorts that also included RAEB and RAEB-t patient samples did not reveal distinctive gene expression or methylation patterns for these malignancies[18 58]. We neither obtained evidence of signatures that could define SF-mutant myeloid malignancies. Nonetheless, we hypothesize that SF-mutant myeloid disorders constitute a biological entity with distinct gene expression and methylation patterns. To address this, we developed an approach towards integrative analysis of the GEP and DMP-datasets to address this hypothesis. Our data point to the existence of two AML/RAEB/RAEB-t clusters each with a different GEP/DMP signature, highly enriched for SF-mutant cases.

MATERIAL AND METHODS

Patients and molecular analyses

Diagnostic bone marrow (BM) or peripheral blood (PB) samples from 344 adults were analyzed; patients were enrolled on HOVON/SAKK protocols -04, -04A, -29, -32, -42 and -43 (available at www.hovon.nl)[202-204]. Patients provided written informed consent in accordance with the Declaration of Helsinki and all trials were approved by the Institutional Review Board of Erasmus University Medical Center. Mutational analyses were carried out as described

previously[24,36,208,209]. Summary of clinical, (cyto)genetical and molecular features of the patients have previously been described[58]. Mutation analyses for the genes *U2AF35, SRSF2* and *SF3B1* were performed by denaturing high-performance liquid chromatography (dHPLC) for all 344 samples in the cohort. Sanger sequencing is subsequently performed on samples with an abnormal dHPLC profile using the primer sets as shown in Table S1. RNA and cDNA synthesis was performed as previously described[18]. Whole Exome Sequencing (WES) has been performed on DNA isolated from RAEB, RAEB-t or AML blasts purified by Ficoll-Hypaque (Nygaard) centrifugation and cryopreserved in aliquots[229]. CD3+ T-cells were expanded from diagnostic bone marrow or peripheral blood specimens and used as controls for WES to determine acquired mutations in AML blasts. Primary cells were seeded in supplemented RPMI (10% FCS/100 U/ml penicillin/streptomycin) at ~1x10^6/ml in a 48 well plate pulsed with 25 µL of CD3/CD28-stimulating Dynabeads (Invitrogen Dynal AS, Oslo, Norway) in the presence of 30 U/mL of rIL-2. Re-stimulation with same concentrations was performed after 7-9 days, and subsequent re-stimulations were applied if deemed necessary based on cell numbers determined by microscopy and flow cytometry. Following magnetic separation of the CD3+ T-cell fraction with MACS CD3 MicroBeads (Miltenyi Biotec, Bergisch Gladbach, Germany) according to the manufacturer's recommendation, CD3+ cell purity was routinely determined >96% by flow cytometry, and, in case of lower purity levels, a second purification was performed.

Pre-processing of gene expression and DNA-methylation profiling

Two high throughput data sets were used in this study: genome wide mRNA expression profiling (GEP) and DNA-methylation profiling (DMP) data for 344 samples. GEP data was generated using Affymetrix HGU133 plus2.0 (Santa Clara, CA, USA),[18,58,62]. Sample processing and quality control were carried out as described previously[18]. Normalization of raw data was processed with Robust Multi-array Average (RMA)[88,230] and probes on the array are remapped to Refseq transcripts using a custom Chip Definition File (CDF)[231]. The custom CDF mapped the original probes to known gene-transcripts for UCSC HG19. DMP-data was generated using the HELP-assay, pre-processed as described previously[58], and annotated using UCSC HG19. GEP and DMP-data are available at the NCBI Gene Expression Omnibus accession numbers GSE14468 and GSE18700 respectively.

Pre-processing and the detection of mutations in whole exome sequence data

RAW-FASTQ files were aligned using Burrows-Wheeler Aligner[101] (BWA) followed by indel realignment using Genome Analysis ToolKit (GATK). The resulting aligned files (e.g. BAM file) were then used to remove PCR duplicates using SAM-tools (Sequence Alignment/Map)[232]. Single nucleotide variant variants (SNV) were called using the unified genotyper of GATK whereas all variants were annotated using Annovar. These annotations were subsequently used to select for nonsynonymous substitutions, stopgain mutations, frameshift insertion or frameshift deletions in the exonic or UTR5 regions that were not reported as a SNP, i.e. by using the Single Nucleotide Polymorphism Database (dbSNP) and Cosmic database. SNVs were also excluded if these were seen in the background, generated by whole exome

sequencing of T-cells of the same patient samples. Coverage and GATK statistics can be found in supplementary Table S2, whereas the frequency of read depth of the aligned loci illustrated in Figure S1.

Statistical Analyses

Differentially expressed and methylated genes for the detected clusters are determined by comparing GEP and DMP-data of each patient sample within the cluster versus patients outside the cluster, using the student T-test. Genes are considered to be differentially expressed or methylated when mRNA or DNA-methylation levels differed with $P \leq 0.001$ after correcting for multiple testing using the Benjamini and Hochberg[233] method (denoted as the false discovery rate; FDR). Patient characteristics among the clusters were compared using the Mann-Whitney-U test (continuous variables) and the Fisher exact test (categorical variables). Outcome measures are assessed using Kaplan-Meier estimates in a univariate analysis. Multivariate analyses were used according the Cox's proportional hazard ratio model. The definition of complete remission (CR) and survival endpoints such as overall survival (OS), event-free survival (EFS), and relapse-free survival (RFS) were based on the recommended consensus criteria[234]. Pathway analysis is performed by utilizing the Molecular Signature Database (MSigDB, v3.0) for the detection of enriched BioCarta pathways, KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) and transcription factor targets. Pathways and/or gene sets are considered statistically significant when the *P-value*, derived from the Hypergeometric test, is less or equal than 0.05 after correcting for multiple testing using FDR. In addition, pathways are derived using Ingenuity Pathway Analysis (Ingenuity® Systems, http://www.ingenuity.com, IPA 8.8) with $P \leq 0.05$.

RESULTS

Hotspot mutations in splice factor genes *SF3B1, U2AF35* and *SRSF2* are more frequent in RAEB/RAEB-t than in AML.

Splice factor mutations (SF-mutations) have been reported to be present in myelodysplastic syndromes (MDS) as well in acute myeloid leukemias (AML). We nucleotide sequenced cDNA of 7 RAEB, 13 RAEB-t and 324 AML patient samples for the eight reported major hotspot mutations[217], i.e. in *SF3B1 (*five hotspots*:* R625L/C; N626D; H662Q/D; K666N/T/E/R; K700E), *U2AF35* (two hotspots: S34F; Q157P) and *SRSF2* (P95H/L/R). In 2/7 RAEB cases we observed mutations in *SRSF2*. Five of the 13 RAEB-t patients carried mutations in *U2AF35* (n=3) or in *SRSF2* (n=2). Since SF-mutations were found in RAEB and RAEB-t samples, and molecular, clinical data revealed no significant differences between the RAEB and RAEB-t groups (Table S3), we combined them for further analysis. In AML we found mutations in *SF3B1* (n=7), *U2AF35* (n=4) and *SRSF2* (n=17) (Table S4). Thus, we observed a much higher frequency of SF-mutations in RAEB(T) (RAEB plus RAEB-t) than in AML (35% vs. 8.6%, $P<0.001$).

Splice factor mutant AML and RAEB(T) are highly similar.

AML patients with *SF3B1, U2AF35 or SRSF2* mutations were older (59 vs. 46 years, $P<0.0001$), showed significantly lower white blood cell counts (24×10^9/L vs. 37×10^9/L, $P=0.029$), presented with lower bone marrow blast percentages

(49% vs. 70%, *P*<0.0001), and had higher erythroblasts percentages (11% vs. 3%, *P*<0.0001) (Table S5), than AMLs without mutations in SF-genes. In contrast, no significant differences in clinical characteristics were observed between RAEB(T) with (n=7) or without (n=13) splice factor gene mutation (Table S6).

Except for the bone marrow blast percentages, (16% in RAEB(T) vs. 49% in AML, *P*<0.0001), the parameter that a priori defines the separation between RAEB, RAEB-t and AML [226], no differences were observed between SF-mutant RAEB(T) (n=7) and SF-mutant AMLs (n=28) (Table S4). Ring sideroblasts were found in bone marrow samples from the SF-mutant RAEB(T) as well as SF-mutant AML patients (Table S4). Thus, SF-mutant RAEB(T) and SF-mutant AML are clinically, cytologically and molecularly similar.

Two distinct AML/RAEB(T) enriched clusters revealed through integrative analysis of gene expression and cytosine methylation profiles.

We next evaluated whether SF-mutant malignancies among the cohort of 324 AML and 20 RAEB(T) patients carried unique combined gene expression (GEP) and DNA-methylation profiles (DMP). We carried out 440 distinct hierarchical clustering analyses, using variable combinations of differentially expressed or differentially cytosine methylated genes (Figure S2A). For each clustering, we addresses whether the grouping of samples was "stable" by computing the significance of the clusters with 1000 multi-scale bootstraps. Subsequently, we computed the silhouette scores[235] from the significant clusters, which does describe how distinctive one cluster is from another one. Using these statistics, we could select the optimal hierarchical clustering without making ad hoc decisions. The criteria used to choose the variable combinations of differentially expressed or differentially cytosine methylated genes and the procedures applied to define which is the most optimal combination of probesets for clustering is explained in the Supplement (Computing the optimal hierarchical clustering). The optimal integrated hierarchical clustering was observed when GEP and DMP were combined using 2168 GEP and 2045 DMP probesets, which resulted in the segregation of 18 clusters (Figure 1 and Figure S2B). For each of the clusters we assessed the enrichment for the currently known molecular and (cyto)genetical abnormalities (Figure 1). AMLs with either inv(16), t(15;17), t(8;21) formed three distinct clusters each (cluster # 1, 9, 10). *CEBPA* double-mutant and *CEBPA*^{silenced} AMLs formed cluster #16 and # 18 respectively. Various other abnormalities, i.e. mutations in *NMP1*, *DNMT3A*, *IDH1* or *IDH2*, *FLT3ITD*, *FLT3TKD* as well as chromosomal abnormalities, 3q, 7q or 11q23 defects are depicted in Figure 1. The distribution of these well characterized AML subsets using GEP or DMP-datasets only are represented in Figure S3. Detailed molecular and cytogenetic data of all AML patients in each cluster are presented in Table S7.

Besides the previously identified AML subgroups, two novel clusters, i.e. #3 (n=25) and #11 (n=19) were apparent. Clusters #3 and #11 are highly enriched for RAEB(T) patients (both *P*<0.0001, Table 1). The unique GEP/DMP signatures that identified clusters #3 and #11 prompted for further study.

Patients in clusters #3 and #11 are enriched for RAEB(T) and AMLs with splice factor gene mutations.

Of the 25 cases in GEP/DMP cluster #3, eight were classified as RAEB(T) (32%; 8/25, *P*<0.0001, Table 1). The cluster was preferentially enriched for splice factor gene hotspot mutations (52%; 13/25, *P*<0.0001), i.e. *SF3B1* (n=2)*, U2AF35* (n=2)*,* and *SRSF2* (n=9) (Figure 1 and Figure 2A, Table 1). Four out of 8 RAEB(T) and 9/17 AML cases carried SF-mutations (Table S8). The patients in cluster #11 were enriched for RAEB(T) (31.6%, 6/19, *P*=0.0003), and hotspot mutations (42.1%, 8/19, *P*<0.0001) as well. The hotspot mutations are detected among *SRSF2* (n=2), *SF3B1* (n=3) *and U2AF35* (n=3) (Figure 1, Figure 2B and Table 1). SF-mutations were seen in 2 out of the 6 RAEB(T) cases, and 6 out of the 13 AML cases (Table S8). In case of using the GEP or DMP-data sets separately, there was some grouping of these (RAEB(T)) patients and SF-mutations, however these were not significantly grouped together for a particular cluster (Figure S3). Thus the grouping of these (RAEB(T)) patients and SF-mutations were only evident when GEP and DMP-data were used in combination.

We considered the possibility that in cases of clusters #3 and #11 that did not carry hotspot SF-mutations other SF-alterations might be present. Whole exome sequencing (WES) was carried out on DNA obtained from non-SF-mutant patients of which material was available, i.e. five samples from cluster #3 and six from cluster #11. We did not find other mutations in any of the 8 splice factor genes previously reported to be frequently mutated. However, three acquired mutations (absent in T-cells from the same patients) were found in other RNA-binding and RNA-splice factor genes. In cluster #11 mutations in *DHX15* (nonsynonymous; patient #6448), *PRPF4B* (Frameshift deletion; patient #2246) and *CELF4* (nonsynonymous; patient #3318) were found (Table S9).

Erythroid phenotype of cluster #11 patient samples.

Morphological analysis of bone marrow samples from patients of clusters #3 and #11 revealed that blast percentages of the two clusters were both significantly lower compared to the other AMLs (*P*<0.0001; 34% vs. 68% (cluster #3) and *P*<0.0001; 31% vs. 68% (cluster #11)) (Table 1). Higher percentages of erythroblasts were found in cluster #11 marrow preparations when compared the other AMLs (*P*<0.0001; 32% vs. 3%), and to cluster #3 (*P*<0.0001; 32% vs. 5%) (Table 1). White blood cell counts (WBC) of cluster #11 cases were significantly reduced in comparison to unselected AMLs (*P*<0.0001; $6x10^9$/; vs. $36x10^9$/L respectively), whereas cluster #3 patients showed WBC counts that were equal to other cases ($31x10^9$/L, Table 1). Thus the two splice factor mutant clusters which are both enriched for RAEB(T) samples show morphological differences for which cluster #11 patients revealed a strong erythroid phenotype (Table 1).

Differentially expressed or hypomethylated genes in cluster #11 patient samples strongly associate with erythroid development.

The signature of 895 differentially expressed and 1180 differentially methylated genes characterized the cases in cluster #11 compared to unselected AMLs. Pathway analysis revealed that the profiles in cluster #11 were highly enriched for gene sets associated with erythroid development and function, e.g. Alpha-Hemoglobin Stabilizing Protein pathway (AHSP)[236,237], Porphyrin metabolism or P53 signaling (Figure 3A). Numerous erythroid genes were found to

be hypomethylated and comparatively overexpressed, such as *GATA1, FECH, ALAS2, AQP1* or *KLF1*. Other erythroid genes were overexpressed with no change in DNA-methylation, such as for *ALAD, UROS, UROD, AHSP or HBD* (Figure 3B and Figure S4). Analysis of transcription factor binding sites, using the differentially expressed and methylated genes revealed significant enrichment for the E2F and GATA1 transcription factor binding sites among these genes (P<0.002 and P<0.001 respectively).

In contrast to cluster #11 AML cases, the 1522 differentially expressed and 74 methylated genes that are associated with cluster #3, lacked the dominant erythroid signature (Figure S5A and B). Thus while both clusters #3 and #11 are enriched for RAEB(T) cases and frequently harbour SF-mutations, cluster #11 cases are specifically associated with a combined myeloid and erythroid phenotype.

Clusters #3 RAEB(T) and AML patients frequently carry RAS mutations.

WES on the small selection of cluster #3 and #11 cases revealed one *KRAS* and 3 *NRAS* mutants among the 6 cases of cluster #3 that were analysed. We applied Sanger sequencing for *NRAS* and *KRAS* among all patients of the two clusters. Ten out of the 25 (40%) patients in cluster #3 carried mutations in *NRAS* (N=9) or *KRAS* (N=1) (P<0.0001, Table 1 and Figure 2). In contrast, no *RAS* mutations were found in any of the cluster #11 cases analyzed (Table 1).

Unfavourable outcome for cluster #3 patients

To verify whether cluster #3 and #11 differed clinically in terms of prognosis, we assessed the overall survival (OS), relapse-free survival (RFS), and event-free survival (EFS). The overall survival for patients in cluster #3 and #11 showed a 5-year OS of 24% (95% CI, 9%-42%) and 41% (95% CI, 20%-62%) respectively (Figure 4 and Figure S6). In an univariate analysis, cluster #3 patients showed significant inferior outcome measures compared to unselected AMLs (OS: P=0.001, Figure 4A, RFS: P=0.014, Figure S6A, EFS: P=0.016, Figure S6B), whereas this was not seen for cluster#11 cases (OS: P=0.425, Figure 4A, RFS: P=0.944, Figure S6A, EFS: P=0.638, Figure S6B). In a multivariate analysis, we could confirm that cluster #3 patients showed a poor treatment response, independent from other relevant covariates with prognostic value (age, white blood cell count (WBC), *FLT3*^{ITD}, *NPM1*^{pos}, *NRAS/KRAS*, and RAEB(T) and high-risk (cyto)genetics), (OS: P=0.042; Figure 4B, RFS: P=0.045; Figure S6C, EFS: P=0.1; Figure S6D). The multivariate analysis did not reach significance for cluster #11 (Figure S6E and F).

DISCUSSION

In this study we evaluated the frequency of SF-mutations in RAEB(T) (RAEB and RAEB-t) and AML patients. We demonstrate that the discrimination between RAEB, RAEB-t and AML, solely based on percentage of blasts is artificial and that SF-mutant AML/RAEB(T) should be viewed as a shared malignancy. According to molecular criteria, i.e. GEP, DMP and RAS mutation analysis two subclasses of AML/RAEB(T) can be recognized. Not all patients, in the two clusters that we identified carried one of the currently well-described hotspot mutations in the splice factor genes *SF3B1*,

U2AF35, and *SRSF2.* We provide data that point to the existence of other mutations in genes encoding RNA-binding/splicing factors. Although our detected mutations in *DHX15*, *PRPF4B* and *CELF4* have not previously been reported in AML, other *DHX* and *PRPF* family members have been found in AML and MDS as reported in The Cancer Genome Atlas (TCGA) http://cancergenome.nih.gov/. Mutations in *DHX15*, *PRPF4B* and *CELF4* are reported the COSMIC database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/). Moreover, Yoshida et al[217] reported mutations in the splice factor gene *PRPF40B*. Together, these observations favour the hypothesis that more splice factor genes may be mutated in the AML/RAEB(T) patients that can be uncovered using GEP and DMP-data sets in a combined manner.

It has previously been demonstrated that distinct molecularly defined AML subtypes cluster using gene expression or DNA-methylation profiles in an unsupervised manner[18,58]. In contrast to AMLs with for instance translocations t(8;21), t(15;17) or with mutations in *CEBPA*, SF-mutant malignancies did not form any unique cluster when GEP or DMP-data were used separately. It was only through integrating these datasets that we were able to identify SF-mutant patients as being distinct from other cases, and consisting of two subgroups. Thus, the hierarchical clustering approach is ideal to find co-expression of genes which is indicative for co-regulation, which means that the clustering may identify genes that have similar functions or are involved in related biological processes. The disadvantage, however, is that clustering only indicates which genes are co-regulated. Thus, it does not lead to a fine resolution of the interaction processes, such as: whether an interaction between two genes is directly or mediated by other genes, or whether a gene is a regulator or regulatee[238]. To gain a more detailed form of the regulatory interactions patterns, thus to address exactly the mechanisms between DNA-methylation and gene expression within a certain cluster, requires a different statistical strategy such as by using Bayesian networks[238], although the experiments demonstrated in Figure 3 may shed some light on this finding. The high contribution of genes that are both differentially hypomethylated and highly expressed in the same patient samples from cluster #11 may explain why these patients clustered so strongly when GEP and DMP-data were studied in an integrated manner (Figure S3A and B). The reason why patients from cluster #3 could only be defined using the combination of GEP and DMP-datasets is unclear, but based on the so called silhouette scores[235] using the bootstrap labels from Pvclust[239] (See Supplement), the hierarchical clustering appeared stable.

We found multiple SF-mutant samples outside clusters #3 and #11. The question is whether, these SF-mutant AMLs are biologically different or whether they were grouped in different clusters due to technical inaccuracies, meaning that they should have been identified as cluster #3 or #11 cases when more sophisticated procedures of gene expression and genome wide cytosine methylation analyses had been applied. Gene chip hybridization experiments that we applied in this study, is nowadays being replaced by RNAseq, a procedure that not only determines gene expression levels, but also discriminates between different splice forms. To study cytosine methylation we applied HELP, an assay that generates “snapshots” of small areas within CpG rich regions. We hypothesize that the combination of RNA-Seq with more sophisticated tools to determine DNA-methylation profiles, will provide

information that will allow us to generate even better combined GEP/DMP signatures. It is possible that SF-mutant cases that were not found in clusters #3 or #11, potentially belong to either of these two clusters but were missed with the currently used methodologies. In any case, our study highlights the potential of combining biological data sets such as gene expression and DNA-methylation profiling data, and shows that with pursuing such a combined approach, leukemia subtypes with a characteristic genotype hidden among the heterogeneity can be uncovered.

Novel cluster #11 was most remarkable for involving MDS and AML, since these samples appeared to share unique erythroid features based on the following findings: 1. Enrichment of pathways associated with erythroid development, when differentially expressed and methylated genes were analysed. 2. Multiple erythroid genes were simultaneously highly expressed and hypomethylated 3. High cytological percentages of erythroblasts. 4. Presence of patient samples with a RAEB or a RAEB-t. 5. Frequent appearance of ring sideroblasts (Table 1). 6. Presence of AML-M6 (erythroid leukemia) cases. Even though morphological classification pointed towards leukemias with strong erythroid developmental defects, the cluster also contained patient samples classified as AML-M0, M1, M2 or M4. These AMLs showed differential expression and hypomethylation of erythroid genes as well, which separated them from other AMLs with the same FAB-class. We conclude that AMLs with defective erythroid development exist more frequently than morphological classification would suggest.

The two AML/RAEB(T) clusters show several differences, among which the high percentages of N-RAS or K-RAS mutations in cluster #3 and not cluster #11 patients. This striking difference between the SF-mutant enriched clusters may explain the much higher white blood cell counts found among cluster #3 samples. It may also clarify the inferior response to treatment of cluster #3 patients. Cluster #3 patients also contain more frequent mutations in *SRSF2*, which has been reported to occur in AMLs that develop upon leukemic transformation from myeloproliferative neoplasms. We hypothesize that the two clusters that we identified represent two different splice factor mutant malignancies, which may embody distinct evolutionary stages of the disease. This would mean that certain cases in cluster #11 may become cluster #3 AMLs in a later phase of the disease, i.e. upon acquiring mutations in N-RAS or K-RAS. No matter the explanation, our data strongly suggest that SF-mutant RAEB(T) and AML constitute a myeloid entity that overrides the separation between AML and MDS and is composed of two subgroups which show overlap but also differ clinically and molecularly.

ACKNOWLEDGEMENTS

The authors are indebted to the colleagues of the bone marrow transplantation group and the molecular diagnostics laboratory of the department of Hematology at Erasmus University Medical Center (Erasmus MC) for storage of samples and molecular analysis of leukemia cells. This research was performed within the framework of CTMM, the Center for Translational Molecular Medicine, project BioCHIP (grant 03O-102). This work was also supported by grants from the National Institutes of Health to R.D. (CA118316); a grant from the Dutch Cancer Society ‘‘Koningin Wilhelmina

Fonds'' to R.D., P.J.M.V., and B.L (EMCR 2006-3522), and a grant from ErasmusMC (MRace) to R.D and E.T. was supported by a research fellow ship from the Dutch Cancer Society ''Koningin Wilhelmina Fonds''.

AUTHOR CONTRIBUTION

Conception and Design: Erdogan Taskesen, Ruud Delwel, and Bob Löwenberg. **Provision of study materials or patients:** Peter Valk, Bob Löwenberg, Ari Melnick and Ruud Delwel. **Collection and assembly of data:** Erdogan Taskesen, Marije Havermans, Kirsten van Lom, Mathijs Sanders, Eric Bindels, Yvette van Norden, Remco Hoogenboezem, Marcel J.T. Reinders, Maria E. Figueroa, Peter Valk, Bob Löwenberg, Ari Melnick and Ruud Delwel. **Data analysis and interpretation:** Erdogan Taskesen, Bob Löwenberg, Ari Melnick and Ruud Delwel. **Manuscript writing:** Erdogan Taskesen, Ruud Delwel, and Bob Löwenberg. **Final approval of manuscript:** All authors have read and approved the final manuscript.

FIGURE LEGENDS

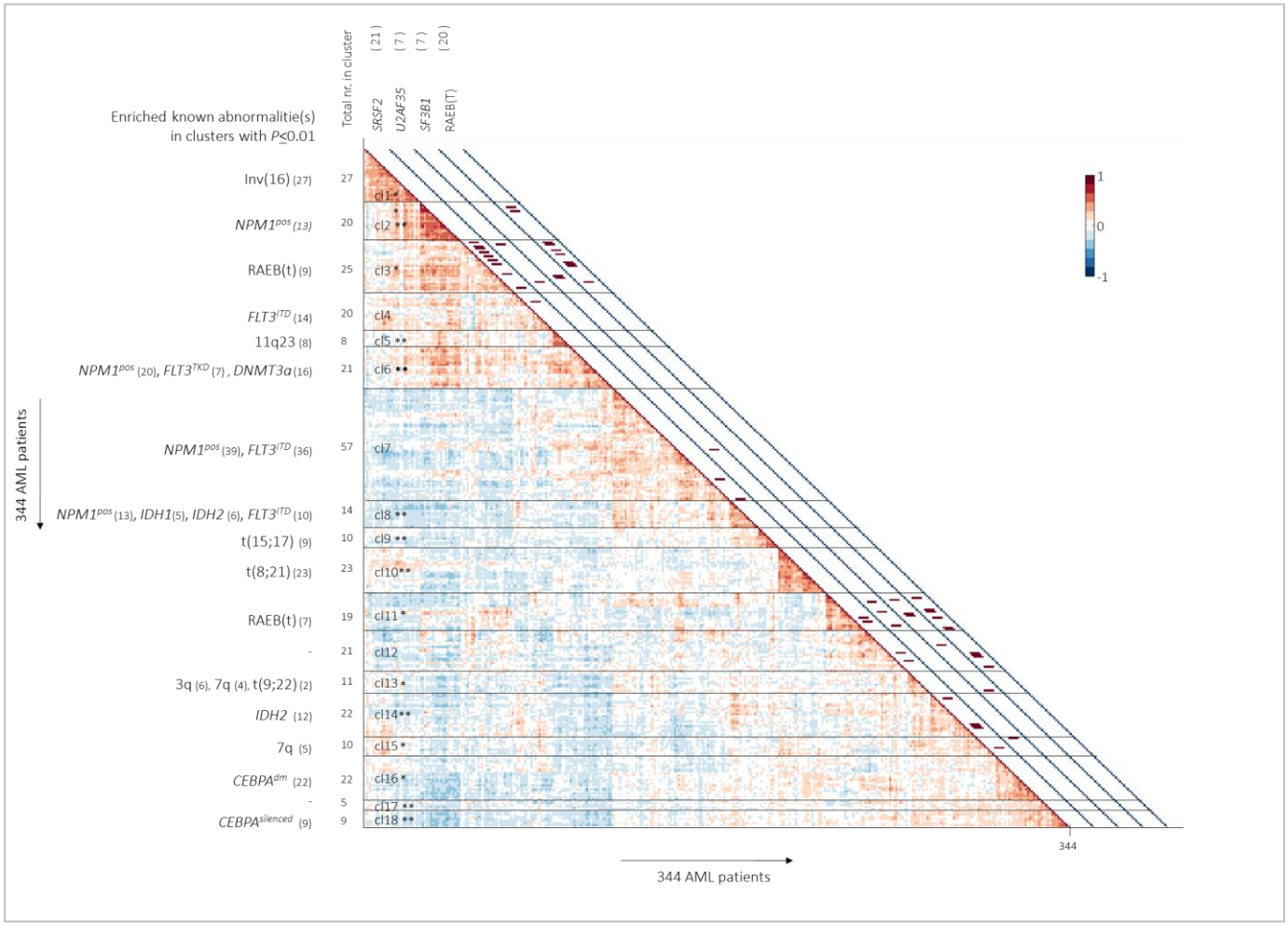


Figure 1. Hierarchical clustering of genetic and epigenetic features segregates AML patients into 18 clusters. Heat map representing pairwise correlations between the 344 AML cases using the gene expression and DNA-methylation profiles of each patient. Ordering of patient samples is based on hierarchical clustering using Pearson correlation and Ward's linkage, which results into clusters of patients that are highly correlated to each other. Colored cells in the heat map depict higher positive (red) or lower negative (blue) correlation, as indicated with the scale bar. Bars in the first four rows along the diagonal of the heat map indicate presence of the splice factor gene hotspot mutations. In the last row it is indicated whether a patient should be considered RAEB(T). Detailed information of each patient in the clusters is shown in Table S7.

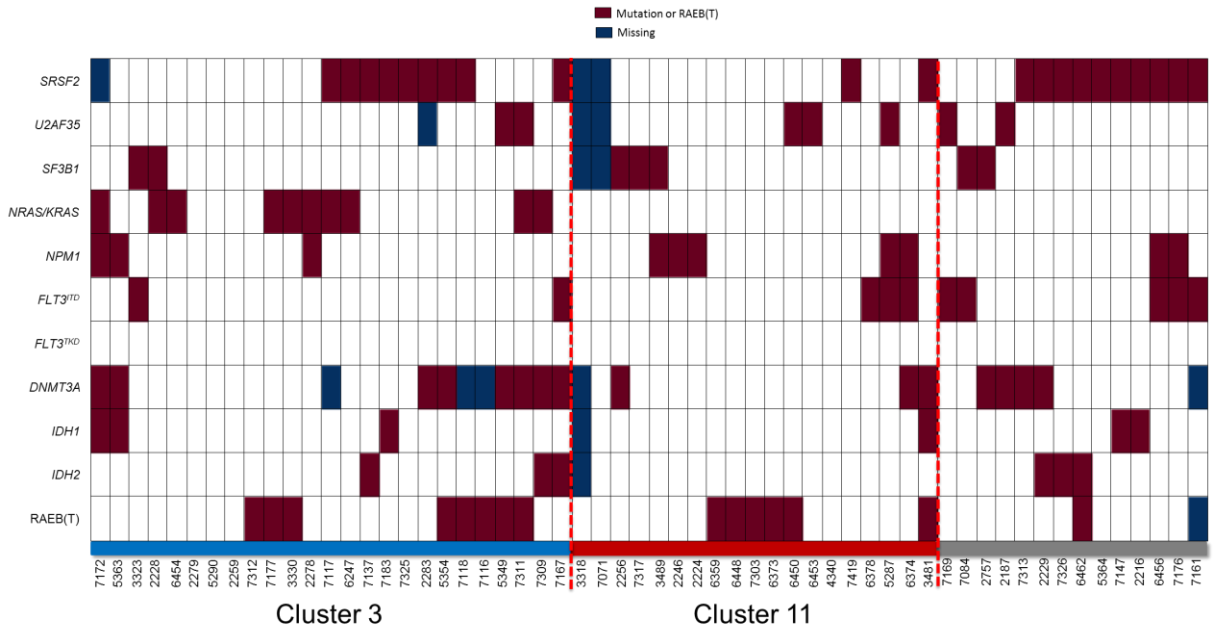


Figure 2. Gene mutations in patients from cluster #3, cluster #11 and splice factor mutations outside these clusters. Columns represent patients from cluster #3, #11 and splice factor mutants outside these clusters. The rows (red) indicates mutations in the genotypes *SRSF2*, *U2AF35*, *SF3B1*, *NRAS/KRAS*, *NPM1*^{mutant}, *FLT3*^{ITD}, *FLT3*^{TKD}, *DNMT3A*, *IDH1* and *IDH2*. Wild-type genotypes are indicated in white and missings in blue. The bottom row indicates the RAEB(T) status in red and missings in blue.

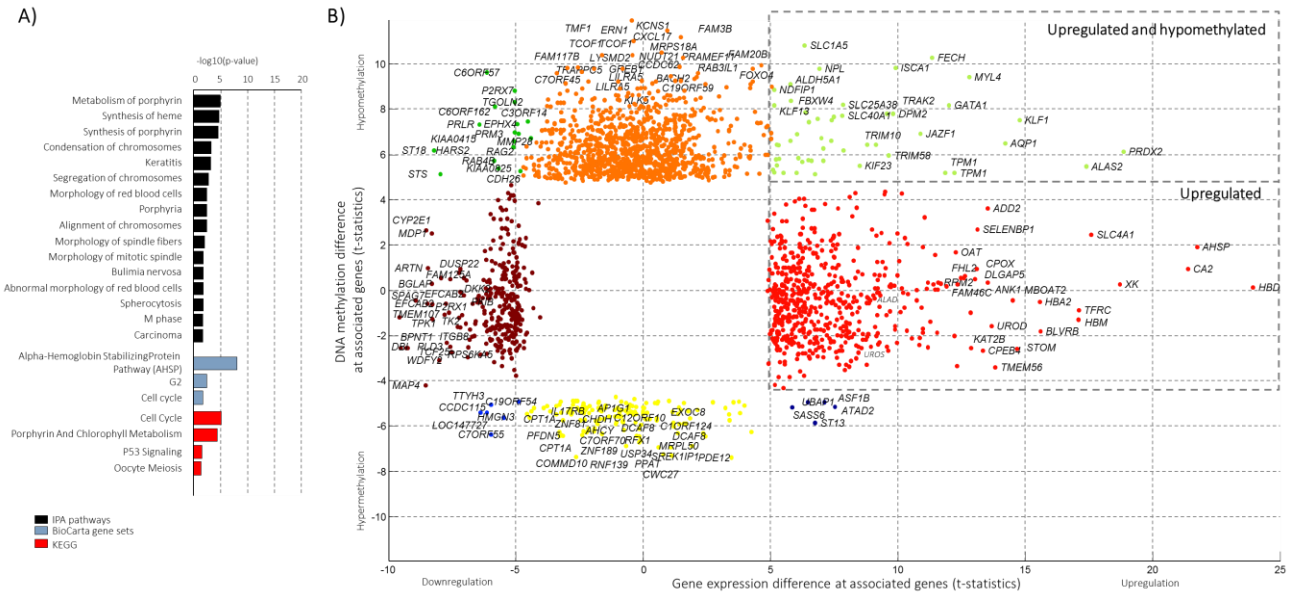
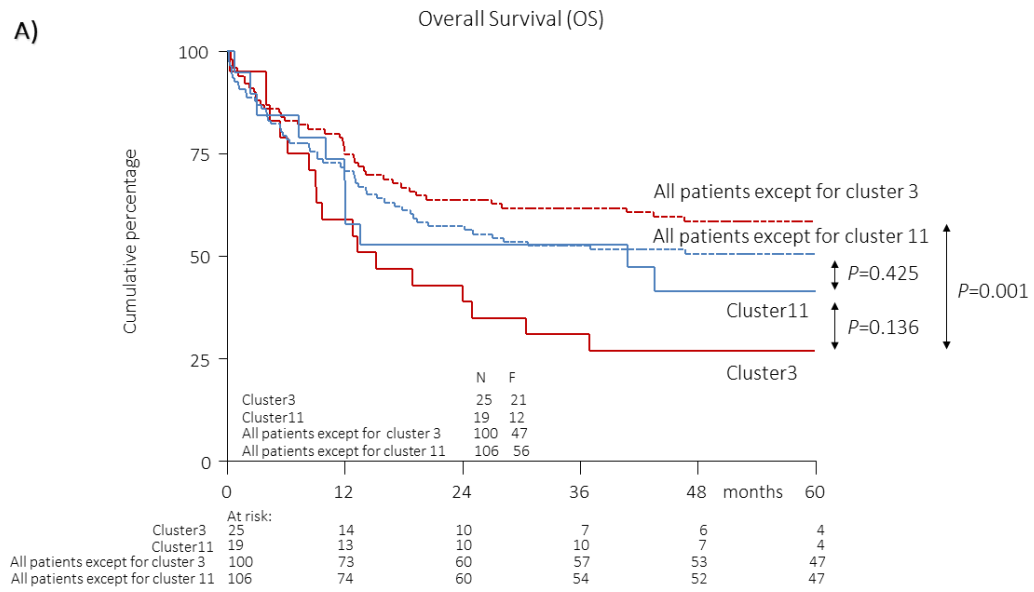


Figure 3. Specific DNA-methylation and gene expression patterns for patient samples from cluster #11 and the enriched pathways. (A) Graphical representation of the top 10 enriched pathways/gene sets based on the differentially expressed and DNA methylated genes for cluster #11. (B) Differential expressed and DNA-methylation genes in patient samples from cluster #11 compared to all other AMLs are indicated with different colored dots. The colors depict the gene expression and DNA-methylation status, i.e. the right upper corner represents genes that are hypomethylated and overexpressed (green dots). Many of these genes encode for proteins involved in erythroid development or function.

A)



B)

Variables	Hazard Ratio (HR)	Std.Err	z	P>|z|	95% Confidence Interval	
					low	high
Cluster 3	1.880	0.584	2.03	0.042*	1.022	3.457
Age	1.022	0.009	2.35	0.019*	1.004	1.041
Log(WBC)	1.003	0.272	0.01	0.991	0.589	1.708
High (cytogenetic) risk group	1.958	0.830	1.58	0.113	0.852	4.496
FLT3^{ITD}	2.419	0.909	2.35	0.019*	1.158	5.052
NPM1^{+}	0.374	0.229	-1.61	0.108	0.113	1.240
N-RAS/K-RAS	1.583	0.521	1.4	0.163	0.831	3.018
RAEB(T)	0.984	0.413	-0.04	0.97	0.432	2.242

Figure 4. Survival analysis for patients in clusters #3 and #11. Kaplan-Meier survival curves and multivariate analysis for overall survival (OS). Multivariate analysis is based on the Cox proportional hazard ratio (HR) model. The included variables into the model are: *NPM1^{mut}* vs. wild-type *NPM1*, *FLT3^{ITD}* vs. no *FLT3^{ITD}*, *NRAS/KRAS^{mut}* vs. wild-type *NRAS/KRAS,* RAEB(T) vs. no RAEB(T), high cytogenetic risk vs. no high cytogenetic risk; age and white blood cell count (WBC) are used as a continues variable. (A) Kaplan-Meier curves for cluster #3 vs. all patients except for cluster #3 patients, cluster #11 vs. all patients except for cluster #11 patients, and cluster #3 vs. cluster #11 patients. (B) Multivariate analysis for cluster #3 patients.

Characteristics	Cluster 3 (n=25)	AML-rest (n=319)	p^1	Cluster 11 (n=19)	AML-rest (n=325)	p^2	p^3
Age, years			0.00026*			0.11	0.17
median	58	47		51	48		
range	18-72	15-77		33-73	15-77		
Missing	0	1		0	1		
Sex			0.41			0.24	1
Male	16 (64%)	171 (54%)		13 (68%)	174 (54%)		
Female	9 (36%)	147 (46%)		6 (32%)	150 (46%)		
Missing	0	1		0	1		
WBC count, (x10^9/L)			0.85			1.1e-06*	7.9e-06*
Median	31	34		6	36		
Range	4.8-128	0.3-274		1.4-33	0.3-274		
Not determined	0	2		0	2		
Platelet count, (x10^9/L)			0.00054*			0.015*	0.67
Median	83	57		80	57		
Range	26-931	7-742		22-374	7-931		
Not determined	0	2		0	2		
Bone marrow blasts (%)			2.5e-06*			7.1e-07*	0.53
Median	34%	68%		31%	68%		
Range	6-88	0-98		8-64	0-98		
Not determined	0	12		0	12		
Normal karyotype	11 (44%)	141 (44.2%)	1	8 (42.1%)	144 (44.3%)	0.82	1
Fab classification							
Fab class M0	0 (0%)	11 (3.45%)	1	0 (0%)	11 (3.38%)	1	1
Fab class M1	0 (0%)	68 (21.3%)	0.0068*	1 (5.26%)	67 (20.6%)	0.14	0.43
Fab class M2	3 (12%)	79 (24.8%)	0.22	7 (36.8%)	75 (23.1%)	0.17	0.074
Fab class M3	0 (0%)	7 (2.19%)	1	0 (0%)	7 (2.15%)	1	1
Fab class M4	8 (32%)	59 (18.5%)	0.12	2 (10.5%)	65 (20%)	0.55	0.15
Fab class M5	4 (16%)	68 (21.3%)	0.62	0 (0%)	72 (22.2%)	0.017*	0.12
Fab class M6	0 (0%)	3 (0.94%)	1	1 (5.26%)	2 (0.615%)	0.16	0.43
Fab class RAEB(T)	8 (32%)	12 (3.76%)	1.7e-05*	6 (31.6%)	14 (4.31%)	0.0003*	1
Not determined	2	3		2	3		
Ring sideroblasts	4 (16%)	10 (3.1%)	0.23	8 (42%)	6 (1.8%)	0.047*	0.088
Not determined	0	286		0	286		
Erythroblasts (%)	5%	3%	0.14	32%	3%	2.6e-09*	2e-05*
Range	1-29	0-59		8-59	0-52		
Not determined	2	138		3	137		
Thrombocytes (%)	64%	62%	0.51	73%	60%	0.33	0.83
Range	12-931	11-413		22-413	11-931		
Not determined	4	294		5	293		
Mutations							
SRSF2	9 (36%)	12 (3.76%)	1.8e-06*	2 (10.5%)	19 (5.85%)	0.29	0.085
U2AF35	2 (8%)	5 (1.57%)	0.083	3 (15.8%)	4 (1.23%)	0.0034*	0.63
SF3B1	2 (8%)	5 (1.57%)	0.087	3 (15.8%)	4 (1.23%)	0.0033*	0.38
NRAS/KRAS	10 (40%)	30 (9.4%)	1.4e-05*	0 (0%)	40 (12.3%)	0.15	0.0023*

Table 1. Patient demographics and clinical characteristics of patients in cluster 3 and 11. Abbreviations: AML-rest, patients that are not in cluster #3 or #11; Number of cases (percentage), median (range) or missing values are depicted were appropriate; WBC count: white blood cell count; Platelet count: number of platelets per 10^9/L; Bone marrow blasts (%): Percentage of Bone marrow blasts; Fab class, morphological classification; M0, minimally differentiated; M1, without maturation; M2, with maturation; M3, hypergranular promyelocytic; M4, myelomonocytic; M5, (a) monoblastic, (b) monocytic; M6, erytroleukemia; RAEB(T), Refractory Anemia with Excess Blasts (in Transformation); Ring sideroblasts: Patient cells showed Yes/ No Ring sideroblasts; Erythroblasts (%): Percentage of Erythroblasts; Thrombo (%): Percentage of thrombocytes; Splice factor mutations: mutations that are detected in the hotspots of gene *SRSF2*, *U2AF35* and *SF3B1*; *NRAS*/*KRAS*: mutations in codon 12,13 or 61; P1 values indicate the comparison of patients in Cluster #3 versus the patients not in cluster #3 (AML-rest); P2 values indicate the comparison of patients in cluster #11 versus the patients not in cluster #11 (AML-rest); P3 values indicate the comparison of patients in cluster

#3 versus the patients in cluster #11; P-values are marked with (*) if lower than 0.05 and are computed using Mann-Whitney-U test (continues variables) and two sided Fisher exact test (categorical variables).

SUPPORTING MATERIAL

Computing the optimal hierarchical clustering

A major disadvantage of a hierarchical clustering approach is the uncertainty of the derived clusters, e.g. the use of different number of features may result in different patient-clusters. We assessed the uncertainty of a hierarchical clustering using Pvclust[239] that builds on multi-scale bootstrap. It computes a bootstrap probability (BP) and an approximately unbiased (AU) *P-value* for each cluster. These *P-value*s indicate how strong a clustering is supported by the data. Clusters with significance level ≤ 0.05 are taken into consideration which indicates that these clusters do not only "seem to exist" but are stable when we perturb the number of observations. In further analyses we used the AU *P-value* for assessment of uncertainty as this is a better approximation than BP *P-value* according to the authors of Pvclust[239].

To select the hierarchical clustering which is best supported by the data, we used the following procedure: *i)* Ranking the feature sets across-patient standard deviation for each data set and selecting an increasing number of probesets using 21 different cut-offs ([0,..,20%]). The selected feature sets for GEP and DMP are then iteratively combined. *ii)* Each feature set (440 in total) is then used for hierarchical cluster analysis with 1000 multi-scale bootstraps, using Ward's linkage and Pearson correlation distance. *iii)* An average silhouette score[235] (relatedness of samples in a cluster and the separation of different clusters) is computed for each significantly observed cluster from Pvclust (Figure S2). *iv)* Subsequently the hierarchical clustering that is best supported by the data is selected.

One should expect that the highest silhouette score from the significant Pvclust clusters should preserve, to some extent, the clusters of currently known abnormalities (*CEBPA*^{silenced}, *CEBPA*^{dm}, inv(16), t(8;21) and t(15;17)). This is in line with our findings as the highest silhouette score from the significant Pvclust clusters also showed a high silhouette score for the currently known clusters. Note that the GEP and DMP-data is mean normalized with unit variance (z-score).

Computing the stability of the detected clusters

The stability for the 18 newly derived clusters is examined using all the derived hierarchical clustering's as described before. The cluster-labels that are determined for each hierarchical clustering are used to determine the average silhouette score for each of the hierarchical clustering (Figure S2). We hypothesized that stable clusters are frequently seen among different hierarchical clustering's. For the optimal hierarchical clustering we detected that ten out of eighteen clusters (# 1, 2, 3, 6, 8, 9, 11, 13, 16, 18) have high silhouette scores [0.5,..,1] based on all other hierarchical clustering's. These include the (cyto)genetically groups such as, inv(16), t(15;17), *CEBPA*^{dm} and *CEBPA*^{silenced}. Five clusters (# 3, 11, 13, 14, 15) varied in silhouette scores [0.4,..,0.5] and, three clusters (# 4, 7 and 12) showed "low" silhouette scores [0,..,0.4]. Based on these average silhouette scores we categorized the clusters into high (n=10 clusters), medium (n=5 clusters) and low (n=3 clusters) stability (Illustrated with **, * and no asterisks respectively).

Figure legends

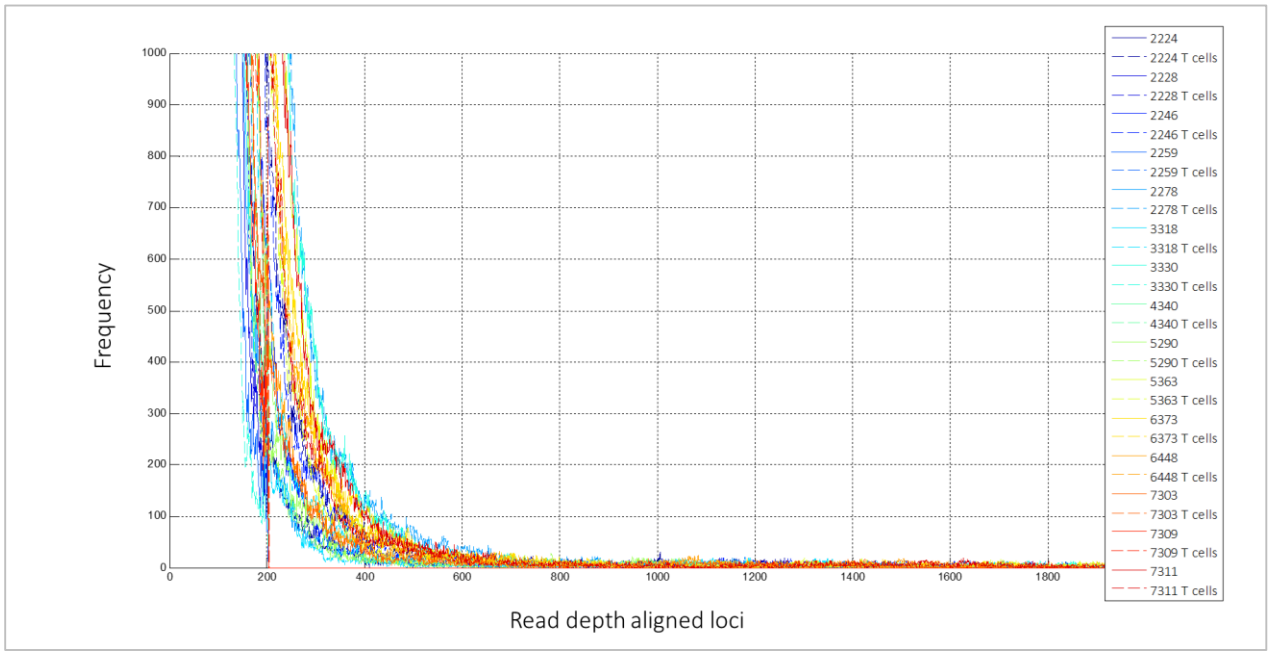


Figure S1. Read depth frequency of the aligned loci. The horizontal axis represents the read depth that is measured for a loci and the vertical axis its frequency. As an example, a loci with read depth of 100 is seen 24517 times for each sample, whereas a read depth of 1000 is seen 4 times for each sample (both are averages among all samples).

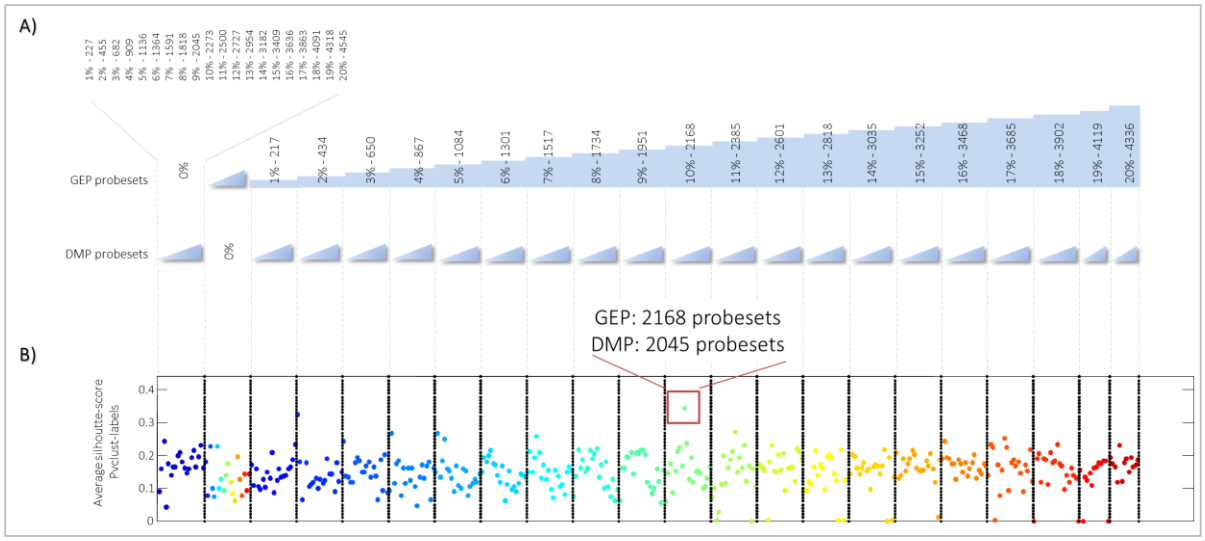


Figure S2. Selection of the most optimal hierarchical clustering. Twenty one different cut-offs are chosen based on patient standard deviation for GEP and separately for DMP. (A) Iteratively combining gene expression and DNA-methylation probesets using one of the 440 combinations: 21x21 minus 1 (zero GEP and zero DMP probesets). (B) The uncertainty of each hierarchical clustering is assessed using the estimated bootstrap labels from Pvclust, and

scored using the silhouette scores. The hierarchical clustering with the highest average silhouette score is best supported by the data and indicated with a red rectangle.

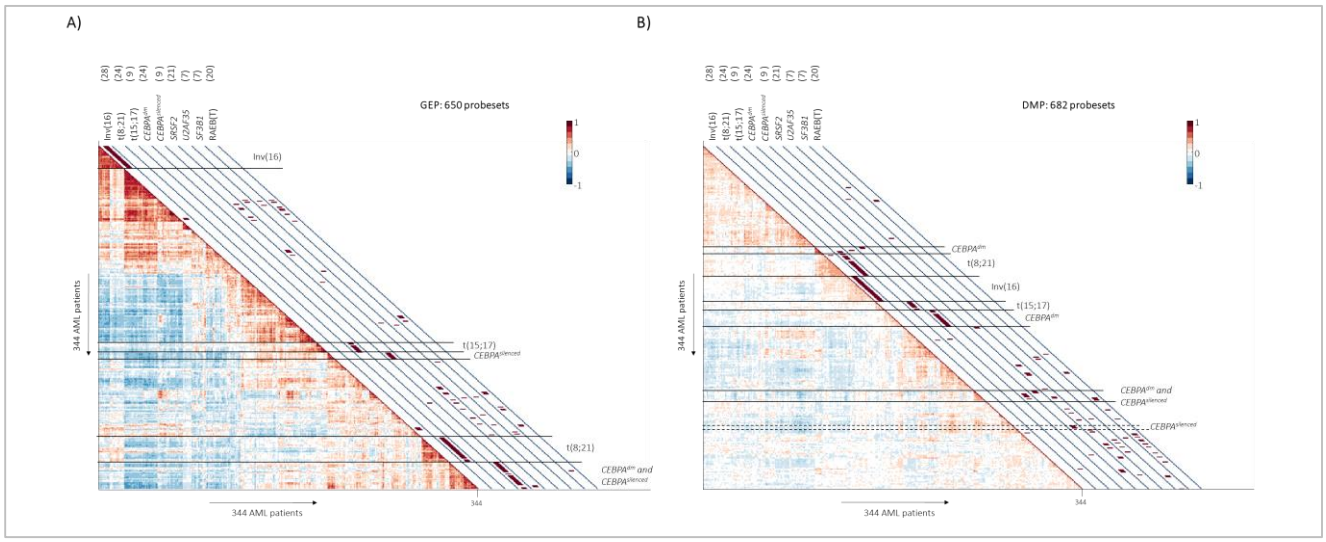


Figure S3. Optimal hierarchical clustering for GEP and DMP separately. Heat map representing pairwise correlations between the 344 AML cases for the gene expression or DNA-methylation profiles. The optimal hierarchical clustering for (A) GEP was based on 650 probesets, (B) DMP was based on 682 probesets. The diagonal illustrates in red the AML subtypes, inv(16), t(8;21), t(15;17), *CEBPA* double-mutants (*CEBPA*^{dm}), *CEBPA*^{silenced}, the splice factor genes *U2AF35*, *SF3B1*, *SRSF2*, and the RAEB(T) cases.

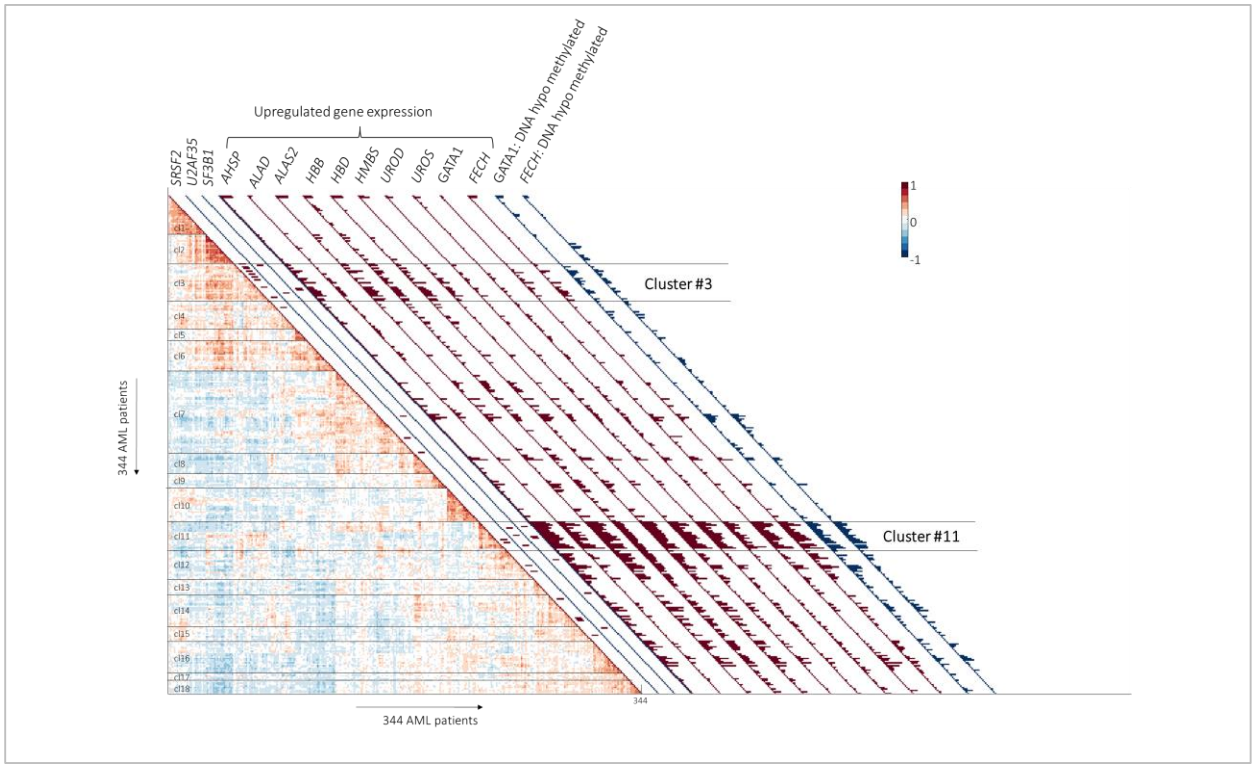


Figure S4. Gene expression and DNA-methylation levels of annotated genes in erythroid development. Heat map representing pairwise correlations between the 344 AML cases using the integrated gene expression or DNA-methylation profiles. The splice factor mutants are indicated with red bars. Gene expression data of the differential regulated genes, involved in erythroid development are mean normalized and depicted on the diagonal with red bars for relative high gene expression levels. Blue bars depict hypomethylation levels.

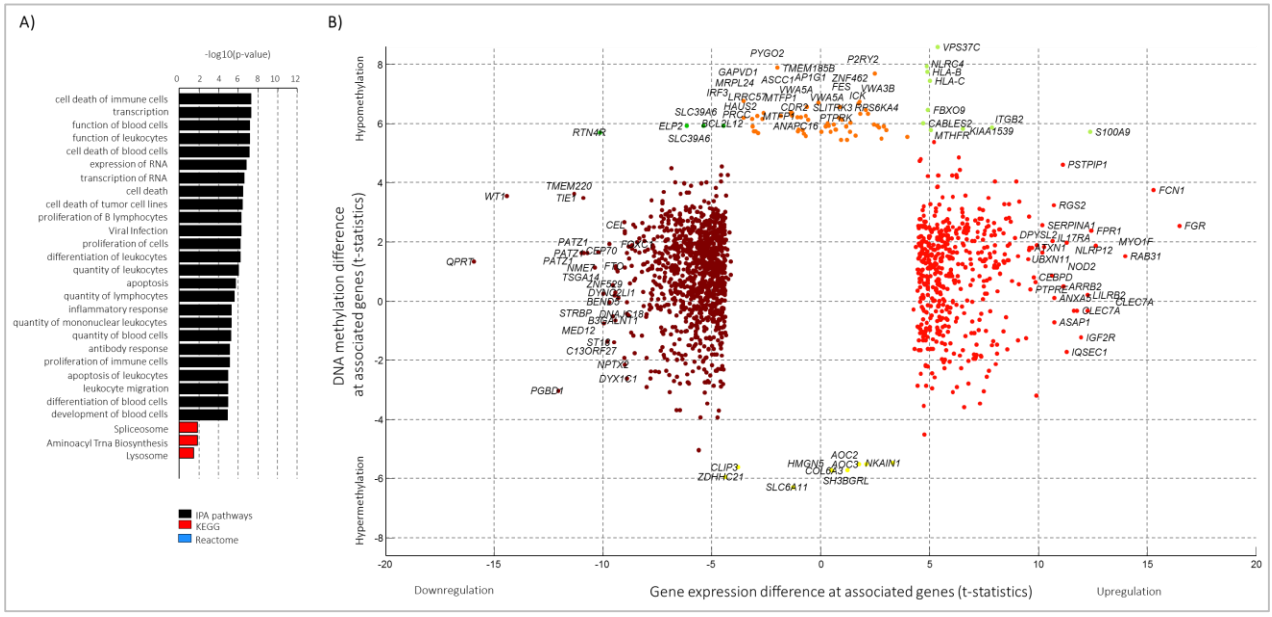
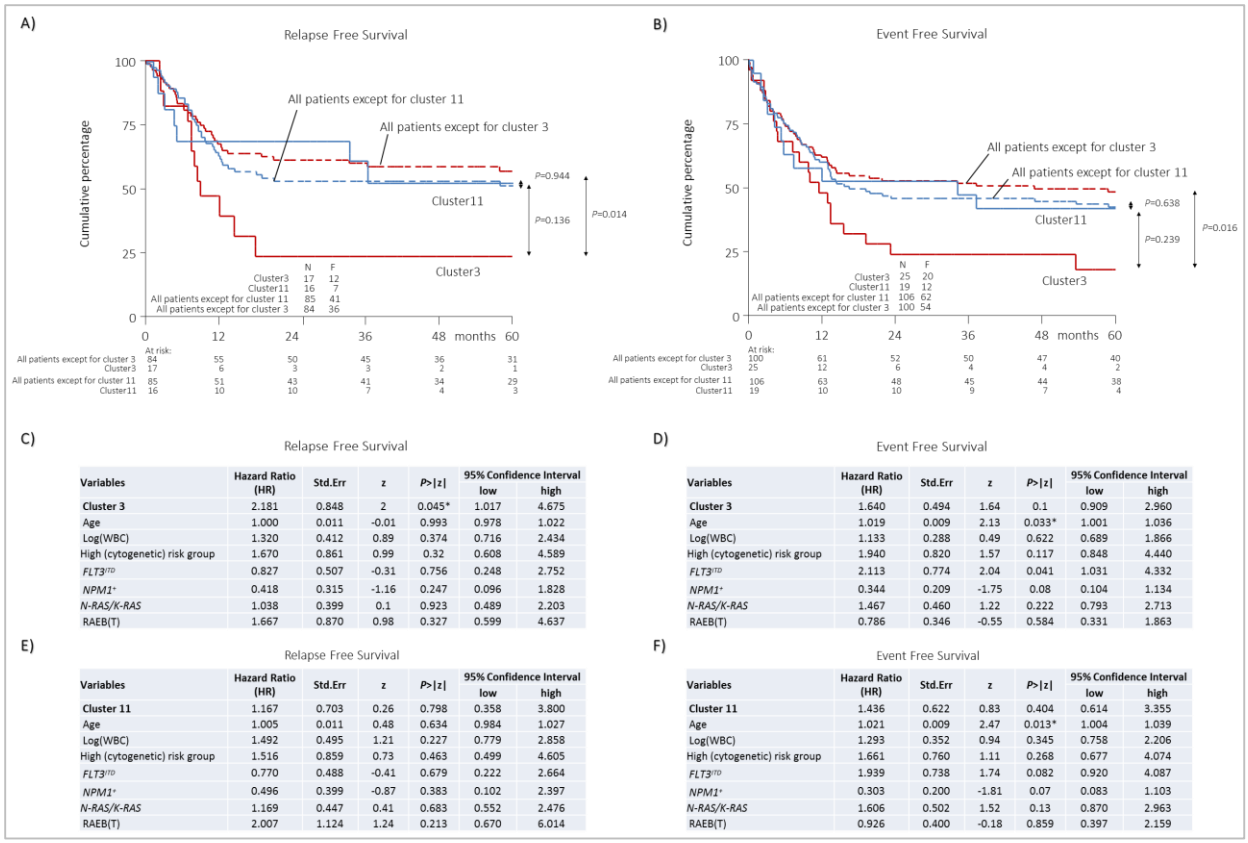


Figure S5. DNA-methylation and gene expression levels for cluster #3 and the enriched pathways. (A) Graphical representation of the top 10 enriched pathways/gene sets based on the differentially expressed and DNA methylated genes for cluster #3. (B) Differential expressed and DNA-methylation genes in patient samples from cluster #3 compared to all other AMLs are indicated with different colored dots. The colors depict the gene expression and DNA-methylation status, i.e. the right upper corner represents genes that are hypomethylated and overexpressed (green dots). Many of these genes encode for proteins involved in erythroid development or function.



C)

Variables	Hazard Ratio (HR)	Std.Err	z	P>|z|	95% Confidence Interval low	high
Cluster 3	2.181	0.848	2	0.045*	1.017	4.675
Age	1.000	0.011	-0.01	0.993	0.978	1.022
Log(WBC)	1.320	0.412	0.89	0.374	0.716	2.434
High (cytogenetic) risk group	1.670	0.861	0.99	0.32	0.608	4.589
FLT3^{ITD}	0.827	0.507	-0.31	0.756	0.248	2.752
NPM1^{+}	0.418	0.315	-1.16	0.247	0.096	1.828
N-RAS/K-RAS	1.038	0.399	0.1	0.923	0.489	2.203
RAEB(T)	1.667	0.870	0.98	0.327	0.599	4.637

D)

Variables	Hazard Ratio (HR)	Std.Err	z	P>|z|	95% Confidence Interval low	high
Cluster 3	1.640	0.494	1.64	0.1	0.909	2.960
Age	1.019	0.009	2.13	0.033*	1.001	1.036
Log(WBC)	1.133	0.288	0.49	0.622	0.689	1.866
High (cytogenetic) risk group	1.940	0.820	1.57	0.117	0.848	4.440
FLT3^{ITD}	2.113	0.774	2.04	0.041	1.031	4.332
NPM1^{+}	0.344	0.209	-1.75	0.08	0.104	1.134
N-RAS/K-RAS	1.467	0.460	1.22	0.222	0.793	2.713
RAEB(T)	0.786	0.346	-0.55	0.584	0.331	1.863

E)

Variables	Hazard Ratio (HR)	Std.Err	z	P>|z|	95% Confidence Interval low	high
Cluster 11	1.167	0.703	0.26	0.798	0.358	3.800
Age	1.005	0.011	0.48	0.634	0.984	1.027
Log(WBC)	1.492	0.495	1.21	0.227	0.779	2.858
High (cytogenetic) risk group	1.516	0.859	0.73	0.463	0.499	4.605
FLT3^{ITD}	0.770	0.488	-0.41	0.679	0.222	2.664
NPM1^{+}	0.496	0.399	-0.87	0.383	0.102	2.397
N-RAS/K-RAS	1.169	0.447	0.41	0.683	0.552	2.476
RAEB(T)	2.007	1.124	1.24	0.213	0.670	6.014

F)

Variables	Hazard Ratio (HR)	Std.Err	z	P>|z|	95% Confidence Interval low	high
Cluster 11	1.436	0.622	0.83	0.404	0.614	3.355
Age	1.021	0.009	2.47	0.013*	1.004	1.039
Log(WBC)	1.293	0.352	0.94	0.345	0.758	2.206
High (cytogenetic) risk group	1.661	0.760	1.11	0.268	0.677	4.074
FLT3^{ITD}	1.939	0.738	1.74	0.082	0.920	4.087
NPM1^{+}	0.303	0.200	-1.81	0.07	0.083	1.103
N-RAS/K-RAS	1.606	0.502	1.52	0.13	0.870	2.963
RAEB(T)	0.926	0.400	-0.18	0.859	0.397	2.159

Figure S6. Relapse-free survival and Event-free survival for patients in clusters #3 and #11. Kaplan-Meier survival curves, and multivariate analysis for (A) relapse-free survival (RFS) and (B) event-free survival (EFS) for cluster #3 vs. all patients except for cluster 3 patients, cluster #11 vs. all patients except for cluster 11 patients, and cluster #3 vs. cluster #11 patients. Multivariate analysis for cluster #3 for (C) RFS, (D) EFS, and for cluster #11 (E) RFS and (F) EFS. The included covariates into the Cox proportional hazard ratio (HR) model are: *NPM1*^{mut} vs. wild-type *NPM1*, *FLT3*^{ITD} vs. no *FLT3*^{ITD}, *NRAS/KRAS*^{mut} vs. wild-type *N/KRAS,* RAEB(T) vs. no RAEB(T), high cytogenetic risk vs. no high cytogenetic risk; age and white blood cell count (WBC) are used as a continues variable.

Gene	Mutation	Primer
U2AF35	Q157 P/R	FW: GTGAGGAAGATGCGGAAAAG
		RV: GGGATCGGGATCTTGATCTAT
U2AF35	34 F/Y	FW: GTATCTGGCCTCCATCTTCG
		RV: TGTTCCTGCATCTCCACATC
SF3B1	N626D	FW: CCCTGGGCATTCCTTCTTTA
	H662Q/D	
	R625L/C	RV: TCGATACCATAAGGAGTTGCTG
	K666N/T/E/R	
	K700E	
SRSF2	P95 H/L/R	FW: GCTGAGGACGCTATGGATG
		RV: ACCGAGATCGAGAACGAGTG

Primer sets Sanger sequencing	
Gene	**Primer**
U2AF35	FW: GTATCTGGCCTCCATCTTCG
	RV: ATCTCTCGACCGCCTCCT
SF3B1	FW: TAGAGTGGAAGGCCGAGAGA
	RV: CTTTGTTTGCCAACTCCACA
SRSF2	FW: CAAGGTGGACAACCTGACCT
	RV: GAGACTTCGAGCGGCTGTA

Table S1. Primer sets dHPLC (WAVE).

Sample	Passed filtered reads	Mapped reads	Mapped reads (%)	R1	R2	Bases covered	Bases aligned	Average coverage per base	Variants called using GATK:	SNV's called using GATK:	substitutions called using GATK:	snp137
2224	34091622	32094531	94.14	15258913	16835618	373056003	1986886090	5.32597	128702	123419	5283	123739
2224_T_cells	47356836	44225068	93.39	21023174	23201894	472512207	2909908293	6.15838	156947	149677	7270	
2228	36372200	34375838	94.51	16189738	18186100	331980963	1563965749	4.71101	118109	114045	4064	115497
2228_T_cells	39905728	37490043	93.95	17537179	19952864	351006415	1713395904	4.88138	123589	119088	4501	
2246	35842032	33590830	93.72	16261848	17328982	415628983	2467425001	5.9366	144932	137818	7114	140237
2246_T_cells	38694732	36390288	94.04	17611303	18778985	451620796	2696995571	5.97181	152047	144855	7192	
2259	34117058	32322925	94.74	15264396	17058529	314595972	1487891607	4.72953	114225	110448	3777	111983
2259_T_cells	48535354	45439523	93.62	21171846	24267677	398600075	2056532498	5.15939	138156	132805	5351	
2278	25438736	24442670	96.08	11985020	12457650	363305533	1870456824	5.14844	125017	119905	5112	121544
2278_T_cells	53288208	50424209	94.63	24400575	26023634	603520411	3791255408	6.2819	186899	178122	8777	
3318	28937768	27416617	94.74	13351220	14065397	351752322	1886557606	5.36331	126724	120930	5794	122998
3318_T_cells	34143650	31973667	93.64	15466801	16506866	397039231	2317539826	5.83705	137582	131213	6369	
3330	52979212	49427482	93.30	23746012	25681470	590082375	3648429392	6.18292	182246	173493	8753	175803
3330_T_cells	32075508	30568871	95.30	14531117	16037754	303970310	1360117971	4.47451	110340	106591	3749	
4340	47247930	44456569	94.09	20832604	23623965	408942685	2013605323	4.92393	149246	143844	5402	145294
4340_T_cells	51454366	48043928	93.37	22316745	25727183	432515914	2163016414	5.00101	150247	144848	5399	
5290	42173302	39338107	93.28	18525214	20812893	431114646	2392225490	5.54893	143426	137360	6066	137176
5290_T_cells	31745666	30204773	95.15	14482603	15722170	376004965	1972052221	5.24475	131462	126021	5441	
5363	55344164	51188451	92.49	24475525	26712926	602900742	3591679671	5.95733	179430	171128	8302	171701
5363_T_cells	43951406	41441614	94.29	20035614	21406000	537334926	3060382687	5.69548	166267	158656	7611	
6373	46915790	44068392	93.93	21243020	22825372	576622910	3241057394	5.62076	171046	163305	7741	165604
6373_T_cells	44088884	41370529	93.83	19959477	21411052	490831326	2970759336	6.05251	154930	147915	7015	
6448	34280072	32414880	94.56	15733795	16681085	414714015	2346944269	5.65919	137129	131142	5987	131892
6448_T_cells	43679746	41084267	94.06	19870521	21213746	488886149	2988629222	6.11314	156717	149587	7130	
7303	34784714	33122410	95.22	16004614	17117796	419375019	2348115345	5.59908	141797	135452	6345	137425
7303_T_cells	49855982	46224692	92.72	21296701	24927991	379789014	2065561416	5.43871	130225	124955	5270	
7309	43975726	40959900	93.14	18972037	21987863	355613368	1832904708	5.15421	124274	119590	4684	121639
7309_T_cells	55736448	51289139	92.02	23420915	27868224	416743495	2256744655	5.41519	137125	131597	5528	
7311	54133686	49991930	92.35	23868057	26123873	597517221	3484262027	5.83123	176468	168593	7875	168639
7311_T_cells	48235262	44651524	92.57	21362775	23288749	537703445	3069161182	5.70791	164881	157590	7291	

Table S2. Coverage and GATK statistics. Abbreviations: Sample: Sample number, Passed filtered reads: Number of reads that are processed by the Illumina software and are present in the Fastq file, Mapped reads: Number of reads mapped on the genome by using BWA, Mapped reads (%): Percentage of mapped reads, R1 and R2: Paired end reads R1 and R2, Bases covered: Number of loci covered on reference genome by the aligned reads (at least one read per locus), Bases aligned: Number of bases aligned on the reference genome, Average coverage: Bases aligned / Bases covered, snp137: the overlap of the variants, SNVs and substitutions that are called using GATK and are present in the snp137 data base.

Characteristics	RAEB (n=7)	RAEB-T (n=13)	P*
Age, years			0.91
median	53	53	
range	20-71	30-72	
Missing	0	0	
Sex			0.63
Male	5 (71%)	11 (85%)	
Female	2 (29%)	2 (15%)	
Missing	0	0	
WBC count, ($x10^9$/L)			0.25
Median	5	11	
Range	1.4-22	2-100	
Not determined	0	0	
Platelet count, ($x10^9$/L)			0.58
Median	135	66	
Range	18-217	13-266	
Not determined	0	0	
Bone marrow blasts (%)			0.78
Median	16%	17%	
Range	12-19	2-28	
Not determined	0	0	
Normal karyotype	2 (28.6%)	4 (30.8%)	1
Erythroblasts	25%	16%	0.78
Range	9-35	1-59	
Not determined	2	1	
Thrombocytes (%)	81%	62%	0.83
Range	19-217	17-266	
Not determined	3	4	
Ring Sideroblasts	2	3	0.78
Not determined	25%	16%	
Mutations			
FLT3 ^{ITD}	0 (0%)	1 (7.69%)	1
FLT3 ^{TKD}	0 (0%)	1 (7.69%)	1
NPM1 $^{+}$	0 (0%)	1 (7.69%)	1
CEBPA double mutation	0 (0%)	1 (7.69%)	1
CEBPA single mutation	1 (14.3%)	1 (7.69%)	1
IDH1	1 (14.3%)	1 (7.69%)	1
IDH2	1 (14.3%)	0 (0%)	0.35
DNMT3A	0 (0%)	4 (30.8%)	0.25
SRSF2	2 (28.6%)	2 (15.4%)	0.6
U2AF35	0 (0%)	3 (23.1%)	0.52
SF3B1	0 (0%)	0 (0%)	1

Table S3. Comparison of patient demographics, clinical and molecular characteristics between: RAEB with RAEB(t). Abbreviations: RAEB(T): Refractory anemia with excess blasts (in transformation); Number of cases (percentage), median (range) or missing values are depicted were appropriate; WBC count: White Blood Cell count; Platelet count: number of platelets per 10^9/L; Bone marrow blasts (%): Percentage of Bone marrow blasts; Normal karyotype: Patient have yes/no normal karyotype; Erythroblasts (%): Percentage of Erythroblasts; Thrombocytes (%): Percentage of thrombocytes; Ring sideroblasts: Patient cells showed Yes/ No Ring sideroblasts; *FLT3ITD*: internal tandem duplication in *FLT3*; *FLT3TKD*: tyrosine kinase domain mutation in FLT3; NPM1: Nucleophosmin 1; *CEPBA* double-mutant: double mutation in *CEBPA*; *CEPBA* single mutant: single mutation in *CEBPA*; *IDH1* or *IDH2*: Isocitrate dehydrogenase 1 or 2; *DNMT3A*: DNA (cytosine-5)-methyltransferase 3A; *SRSF2*, *U2AF35* and *SF3B1* are splice factor mutations that are detected in the hotspots; P-values indicate the comparison between the two groups. P-values are marked with (*) if lower than 0.05 and are computed using Mann-Whitney-U test (continues variables) and two sided Fisher exact test (categorical variables).

Characteristics	RAEB(T) with splice factor mutations (n=7)	AMLs with splice factor mutations (n=28)	P*
Age, years			0.37
median	65	59	
range	46-72	37-77	
Missing	0	0	
Sex			0.39
Male	6 (86%)	17 (61%)	
Female	1 (14%)	11 (39%)	
Missing	0	0	
WBC count, (x10^9/L)			0.38
Median	5	24	
Range	2-100	2.1-109	
Not determined	0	0	
Platelet count, (x10^9/L)			1
Median	65%	65%	
Range	35-135	10-931	
Not determined	0	0	
Bone marrow blasts (%)			0.00045*
Median	16%	49%	
Range	9-25	6-93	
Not determined	0	3	
Normal karyotype	2 (28.6%)	11 (39.3%)	0.67
Fab class			
M0	0 (0%)	1 (3.57%)	1
M1	0 (0%)	4 (14.3%)	0.56
M2	0 (0%)	8 (28.6%)	0.17
M3	0 (0%)	0 (0%)	1
M4	0 (0%)	7 (25%)	0.3
M5	0 (0%)	4 (14.3%)	0.56
M6	0 (0%)	0 (0%)	1
RAEB(T)	7 (100%)	0 (0%)	1.5e-07*
Not determined	0	3	
Erythroblasts (%)	29%	11%	0.42
Range	1-59	1-52	
Not determined	0	7	
Thrombocytes (%)	46%	60%	0.52
Range	17-127	11-931	
Not determined	1	6	
Ring Sideroblasts	1 (14.3)	7 (25)	1
Not determined	0	0	
Mutations			
FLT3^{ITD}	0 (0%)	8 (28.6%)	0.17
FLT3^{TKD}	0 (0%)	0 (0%)	1
NPM1$^+$	0 (0%)	4 (14.3%)	0.56
CEBPA double mutation	0 (0%)	1 (3.57%)	1
CEBPA single mutation	1 (14.3%)	1 (3.57%)	0.36
IDH1	1 (14.3%)	3 (10.7%)	1
IDH2	1 (14.3%)	4 (14.3%)	1
DNMT3A	4 (57.1%)	7 (25%)	0.15
SRSF2	4 (57.1%)	17 (60.7%)	1
U2AF35	3 (42.9%)	4 (14.3%)	0.13
SF3B1	0 (0%)	7 (25%)	0.3

Table S4. Comparison of patient demographics, clinical and molecular characteristics between. RAEB(T) with Splice factor mutations versus AMLs with Splice factor mutations: Abbreviations: RAEB(T): Refractory anemia with excess blasts (in transformation). With splice factor mutations: mutations that are detected in the hotspots of gene *SRSF2*, *U2AF35* and *SF3B1*; Number of cases (percentage), median (range) or missing values are depicted were appropriate; WBC count: white blood cell count; Platelet count: number of platelets per 10^9/L; Bone marrow blasts (%): Percentage of Bone marrow blasts; Normal karyotype: Patient have Yes/No normal karyotype; Fab class, morphological classification; M0, minimally differentiated; M1, without maturation; M2, with maturation; M3, hypergranular promyelocytic; M4, myelomonocytic; M5, (a) monoblastic, (b) monocytic; M6, erytroleukemia; Erythroblasts (%): Percentage of Erythroblasts; Thrombocytes (%): Percentage of thrombocytes; Ring sideroblasts: Patient cells showed Yes/No Ring sideroblasts; *FLT3ITD*: internal tandem duplication in *FLT3*; *FLT3TKD*: tyrosine kinase domain mutation in *FLT3*; *NPM1*: Nucleophosmin 1; *CEPBA* double-mutant: double mutation in CEBPA; CEPBA single mutant: single mutation in *CEBPA*; *IDH1* or *IDH2*: Isocitrate dehydrogenase 1 or 2; *DNMT3A*: DNA (cytosine-5)-methyltransferase 3A; *SRSF2*, *U2AF35* and *SF3B1* are splice factor mutations that are detected in the hotspots; P-values indicate the

comparison between the two groups. P-values are marked with (*) if lower than 0.05 and are computed using Mann-Whitney-U test (continues variables) and two sided Fisher exact test (categorical variables).

Characteristics	AML with splice factor mutations (n=28)	AMLs without splice factor mutations (n=283)	P*
Age, years			6.6e-07*
median	59	46	
range	37-77	15-77	
Missing	0	1	
Sex			0.43
Male	17 (61%)	144 (51%)	
Female	11 (39%)	138 (49%)	
Missing	0	1	
WBC count, ($x10^9$/L)			0.029*
Median	24	37	
Range	2.1-109	0.3-274	
Not determined	0	2	
Platelet count, ($x10^9$/L)			0.17
Median	65%	56%	
Range	10-931	7-742	
Not determined	0	2	
Bone marrow blasts (%)			0.00092*
Median	49%	70%	
Range	6-93	0-98	
Not determined	3	8	
Normal karyotype	11 (39.3%)	131 (46.3%)	0.84
Fab class			
M0	1 (3.57%)	10 (3.53%)	1
M1	4 (14.3%)	60 (21.2%)	0.47
M2	8 (28.6%)	69 (24.4%)	0.65
M3	0 (0%)	7 (2.47%)	1
M4	7 (25%)	59 (20.8%)	0.63
M5	4 (14.3%)	65 (23%)	0.35
M6	0 (0%)	3 (1.06%)	1
RAEB(T)	0 (0%)	0 (0%)	1
Not determined	3	2	
Erythroblasts (%)			0.00064*
Median	11%	3%	
Range	1-52	0-54	
Not determined	7	123	
Mutations			
FLT3 ^{ITD}	8 (28.6%)	84 (29.7%)	1
FLT3 ^{TKD}	0 (0%)	38 (13.4%)	0.033*
NPM1 $^{+}$	4 (14.3%)	97 (34.3%)	0.034*
CEBPA double mutation	1 (3.57%)	22 (7.77%)	0.71
CEBPA single mutation	1 (3.57%)	8 (2.83%)	0.58
IDH1	3 (10.7%)	21 (7.42%)	0.47
IDH2	4 (14.3%)	28 (9.89%)	0.51
DNMT3A	7 (25%)	67 (23.7%)	0.81
SRSF2	17 (60.7%)	0 (0%)	5.0e-21*
U2AF35	4 (14.3%)	0 (0%)	4.7e-05*
SF3B1	7 (25%)	0 (0%)	2.3e-08*

Table S5. Comparison of patient demographics, clinical and molecular characteristics between. AML with splice factor mutations versus AMLs without splice factor mutations: Abbreviations: With splice factor mutations: mutations that are detected in the hotspots of gene *SRSF2*, *U2AF35* and *SF3B1*; Number of cases (percentage), median (range) or missing values are depicted were appropriate; WBC count: white blood cell count; Platelet count: number of platelets per 10^9/L; Bone marrow blasts (%): Percentage of Bone marrow blasts; Normal karyotype: Patient have Yes/No normal karyotype; Fab class, morphological classification; M0, minimally differentiated; M1, without maturation; M2, with maturation; M3, hypergranular promyelocytic; M4, myelomonocytic; M5, (a) monoblastic, (b) monocytic; M6,

erytroleukemia; RAEB(T), Refractory Anemia with Excess Blasts (in Transformation); Erythroblasts (%): Percentage of Erythroblasts; *FLT3ITD*: internal tandem duplication in *FLT3*; *FLT3TKD*: tyrosine kinase domain mutation in *FLT3*; *NPM1*: Nucleophosmin 1; *CEPBA* double-mutant: double mutation in CEBPA; *CEPBA* single mutant: single mutation in CEBPA; *IDH1* or *IDH2*: Isocitrate dehydrogenase 1 or 2; *DNMT3A*: DNA (cytosine-5)-methyltransferase 3A; *SRSF2*, *U2AF35* and *SF3B1* are splice factor mutations that are detected in the hotspots; P-values indicate the comparison between the two groups. P-values are marked with (*) if lower than 0.05 and are computed using Mann-Whitney-U test (continues variables) and two sided Fisher exact test (categorical variables).

Characteristics	RAEB(T) with splice factor mutations (n=7)	RAEB(T) without splice factor mutations (n=13)	P*
Age, years			0.05
median	65	52	
range	46-72	20-71	
Missing	0	0	
Sex			1
Male	6 (86%)	9 (75%)	
Female	1 (14%)	3 (25%)	
Missing	0	0	
WBC count, ($x10^9$/L)			0.95
Median	5	10	
Range	2-100	1.4-22	
Not determined	0	0	
Platelet count, ($x10^9$/L)			0.97
Median	65	67	
Range	35-135	13-266	
Not determined	0	0	
Bone marrow blasts (%)			0.89
Median	16%	17%	
Range	9-25	8-28	
Not determined	0	0	
Normal karyotype	2 (28.6%)	4 (33.3%)	1
Erythroblasts (%)			0.86
Median	29%	22%	
Range	1-59	5-54	
Not determined	0	3	
Thrombocytes (%)	46	65	0.37
Range	17-127	19-266	
Not determined	1	5	
Ring Sideroblasts	1 (14.3)	3 (25)	0.57
Not determined	0	4	
Muations			
FLT3 ^{ITD}	0 (0%)	1 (8.33%)	1
FLT3 ^{TKD}	0 (0%)	1 (8.33%)	1
NPM1 $^{+}$	0 (0%)	1 (8.33%)	1
CEBPA double mutation	0 (0%)	1 (8.33%)	1
CEBPA single mutation	1 (14.3%)	1 (8.33%)	1
IDH1	1 (14.3%)	1 (8.33%)	1
IDH2	1 (14.3%)	0 (0%)	0.37
DNMT3A	4 (57.1%)	0 (0%)	0.0063*
SRSF2	4 (57.1%)	0 (0%)	0.009*
U2AF35	3 (42.9%)	0 (0%)	0.036*
SF3B1	0 (0%)	0 (0%)	1

Table S6. Comparison patient demographics, clinical and molecular characteristics between. RAEB(T) with Splice factor mutations versus RAEB(T) without Splice factor mutations: Abbreviations: With splice factor mutations: mutations

that are detected in the hotspots of gene *SRSF2*, *U2AF35* and *SF3B1*; Number of cases (percentage), median (range) or missing values are depicted were appropriate; WBC count: white blood cell count; Platelet count: number of platelets per 10^9/L; Bone marrow blasts (%): Percentage of Bone marrow blasts; Normal karyotype: Patient have Yes/No normal karyotype; Fab class, morphological classification; M0, minimally differentiated; M1, without maturation; M2, with maturation; M3, hypergranular promyelocytic; M4, myelomonocytic; M5, (a) monoblastic, (b) monocytic; M6, erytroleukemia; RAEB(T), Refractory Anemia with Excess Blasts (in Transformation); Erythroblasts (%): Percentage of Erythroblasts; Thrombocytes (%): Percentage of thrombocytes; Ring sideroblasts: Patient cells showed Yes/No Ring sideroblasts; *FLT3ITD*: internal tandem duplication in *FLT3*; *FLT3TKD*: tyrosine kinase domain mutation in *FLT3*; *NPM1*: Nucleophosmin 1; *CEPBA* double-mutant: double mutation in CEBPA; CEPBA single mutant: single mutation in *CEBPA*; IDH1 or IDH2: Isocitrate dehydrogenase 1 or 2; *DNMT3A*: DNA (cytosine-5)-methyltransferase 3A; *SRSF2*, *U2AF35* and *SF3B1* are splice factor mutations that are detected in the hotspots; P-values indicate the comparison between the two groups. P-values are marked with (*) if lower than 0.05 and are computed using Mann-Whitney-U test (continues variables) and two sided Fisher exact test (categorical variables).

<Table S7 is not included>

Table S7. Characteristics Cluster #1-18. Patient: patient number. Cluster: cluster number. FAB: FAB subtype of AML. Real-time PCR for *CBFA-MYH11*, *PML-RARα* and *AML1-ETO*. *FLT3ITD*: internal tandem duplication in *FLT3*. *FLT3TKD*: tyrosine kinase domain mutation in FLT3. *N-RAS* or *K-RAS*: mutation in codon 12,13 or 61. *NPM1*: Nucleophosmin 1. *DNMT3A*: DNA (cytosine-5)-methyltransferase 3A. *IDH1* or *IDH2*: Isocitrate dehydrogenase 1 or 2. *CEPBA* double-mutant: double mutation in *CEBPA*. *CEPBA* single mutant: single mutation in *CEBPA*. Karyotype: t(15;17), t(8;21), inv(16)/t(16;16),+8,+11,+21,-5(q),-7(q),t(9;22),3q abnormalities, 11q23 abnormalities (translocation/self-fusion (sMLL)), complex(abnormalities involved) (>3abnormalities), and normal karyotype (NN) are indicated. ND: Not Determined.

Cluster	Patient ID	RAEB(T)	WBC count ($x10^9$/L)	Platelet count ($x10^9$/L)	Bone marrow blasts (%)	Ring sideroblasts	Erythroblasts (%)	SF-mutant	Thrombocytes (%)
	5349	RAEB(T)	31.0	86	25	-	1	Yes	57
	5354	RAEB(T)	54.0	54	20	-	2	Yes	18
	7311	RAEB(T)	99.5	46	9	-	5	Yes	17
	7116	RAEB(T)	8.4	266	24	-	5	-	266
	7118	RAEB	4.8	35	13	-	29	Yes	35
	3330	RAEB	22.0	145	41	-	ND	-	ND
	7177	RAEB	19.2	26	19	-	25	-	19
	7312	RAEB	13.0	217	17	Yes	15	-	217
	7309	-	35.0	70	22	-	1	-	21
	7117	-	109.0	64	6	-	10	Yes	64
	2228	-	38.0	931	34	-	16	Yes	931
	2259	-	57.1	74	49	-	3	-	75
3	2278	-	28.0	78	46	Yes	23	-	12
	2279	-	23.0	104	59	-	ND	-	ND
	2283	-	47.4	64	49	-	18	Yes	27
	3323	-	23.0	83	42	Yes	11	Yes	64
	5290	-	8.4	181	49	-	16	-	130
	5363	-	54.0	191	31	-	6	-	117
	6247	-	11.2	115	34	-	1	Yes	ND
	6454	-	43.0	42	63	-	1	-	93
	7137	-	37.0	283	74	-	5	Yes	283
	7167	-	60.7	57	52	-	16	Yes	57
	7172	-	128.0	211	88	-	2	ND	211
	7183	-	12.8	195	66	-	2	Yes	145
	7325	-	26.0	65	29	Yes	1	Yes	ND
	3481	RAEB(T)	2.0	65	20	-	37	Yes	ND
	6450	RAEB(T)	3.9	134	14	Yes	59	Yes	127
	6359	RAEB(T)	7.1	57	10	Yes	44	-	65
	6373	RAEB(T)	13.0	67	17	Yes	ND	-	92
	6448	RAEB(T)	11.0	66	8	-	34	-	62
	7303	RAEB(T)	5.8	23	11	-	54	-	23
	7317	-	16.5	135	17	Yes	52	Yes	46
	2224	-	1.4	22	24	-	54	-	22
	2246	-	2.4	198	54	Yes	41	-	196
11	2256	-	4.7	152	31	Yes	25	Yes	169
	3318	-	6.6	87	35	Yes	19	ND	47
	3489	-	5.0	374	64	-	8	Yes	413
	4340	-	14.3	64	33	-	9	-	86
	5287	-	13.5	80	39	-	18	Yes	80
	6374	-	33.0	71	63	-	21	-	ND
	6378	-	11.8	160	61	-	18	-	ND
	6453	-	2.1	87	47	Yes	30	Yes	57
	7071	-	2.0	125	31	-	ND	ND	ND
	7419	-	5.0	51	35	-	ND	Yes	ND

Table S8. Bone marrow blast, Ring-sideroblasts, Erythroblast and Thrombo characteristics Cluster #3 and 11. Abbreviations: Cluster: Patients are detected in cluster 3 or 11; Patient ID: Patient identification number; ; RAEB(T), Refractory Anemia with Excess Blasts (in Transformation); WBC count: white blood cell count; Platelet count: number of platelets per 10^9/L; Bone marrow blasts (%): Percentage of Bone marrow blasts; Ring sideroblasts: Patient cells showed Yes/No Ring sideroblasts; Erythroblasts (%): Percentage of Erythroblasts; SF-mutants: Splice factor mutation in *SRSF2*, *U2AF35* or *SF3B1*; Thrombocytes (%): Percentage of thrombocytes.

Patient	Genomic location	Reference	Altered	Variation situated	Gene affected	Gene effect	Genotype
6448	chr4: 24572314-24572314	G	C	exonic	*DHX15*	nonsynonymous SNV	DHX15:uc003gqx.3:exon3:c.C664G:p.R222G
2246	chr6: 4032794-4032794	G	_	exonic	*PRPF4B*	frameshift deletion	PRPF4B:uc003mvv.3:exon2:c.1043delG:p.R348fs
2246	chr6: 4032796-4032796	A	T	exonic	*PRPF4B*	nonsynonymous SNV	PRPF4B:uc003mvv.3:exon2:c.A1045T:p.S349C
3318	chr18: 34901838-34901838	G	A	exonic	*CELF4*	nonsynonymous SNV	CELF4:uc002laf.2:exon3:c.C364T:p.R122W

Table S9. Newly identified putative splice factor mutations. Patient: patient number. Reference: Reference human genome 19 sequence. Altered: Measured variation in DNA sequence. Variation situated: genomic-location of the

detected variation in the DNA sequence. Gene affected: The gene that is affected by the variation the DNA sequence. Gene effect: nonsynonymous SNV (Single Nucleotide Variation) alters the amino acid sequence of a protein, Frameshift deletion: a number of nucleotides that is deleted in the DNA sequence. Genotype: Characteristics of the detected variation.

CHAPTER 10

General Discussion

General Discussion

ABSTRACT

This thesis is divided into four sections in which we aim to develop and apply statistical approaches to understand the meaning of genome wide molecular data determined in the cells of patients with Acute Myeloid Leukemia. Central in the studies is one group of patients that have abnormalities in the gene called *CEBPA*. Although we addressed very specific biological questions regarding *CEBPA* in AML, the statistical approaches that are presented can also be used for other groups of patients. In this Chapter, we discuss potential future research directions and the results in an integrated manner.

IDENTIFICATION OF POTENTIAL FUNCTIONAL REGIONS IN THE GENOME

In Chapter 2, 3 and 4 we describe the development, implementation and application of HAT(SEQ) (Hypergeometric Analysis of Tiling-arrays and Sequence data), a method to detect potential functional regions in the genome defined with chromatin immunoprecipitation on chip (ChIP-on-chip) or by massively parallel DNA sequencing (ChIP-Seq) data. Together with HATSEQ, multiple other methods[74,147,151,240-242] have been developed to analyze tiling-array and NGS data with the purpose to define potential functional regions (regions-of-interest, ROI). The various methods use different statistical approaches and therefore, if these methods are used to analyze the same dataset, differences in the results can be found. An obvious question is therefore: "*What is the best method to use?*".

The most common approach that is used to test the performance of different methodologies is by overlaying the detected regions. It has been shown that different methodologies do show significant overlap in detected regions[95]. It is very likely that these similarly detected regions are represented by genomic regions with high signals (read depth or intensity values). The differences in detected regions between various methodologies are likely the regions with subtle changes (low read depth or intensity values) or an “unexpected” size of the region. A reason why one methodology may detect a particular region whereas another does not, may be because methodologies can be designed and optimized for the analysis of one type of tiling-array or ChIP-Seq application; the expected signal and/or size of the binding region is then modeled. Unfortunately, there is no golden standard that defines such parameters of candidate regions as it depends on the biological experiments (e.g. protein properties). Thus, overlaying the detected regions using different methodologies will not per definition define true binding regions but only emphasizes the similarity of detected regions between the methodologies.

In general, the most straightforward way to measure the performance of a methodology is by validating each detected candidate region by directed PCR and Sanger sequencing[243]. However, this process is expensive and time consuming when even tens of candidate regions need to be validated. In general, hundreds or thousands of candidate regions can be detected in a single experiment, and consequently, PCR-Sanger sequencing is not an option to validate all ROIs.

An approach to test the performance of a method can be by using in silico data sets[244-248] (e.g. simulated data sets). In these simulated data sets, the "true binding regions" and background-noise-signal (technical variation) is included. However, a disadvantage of in silico data sets is the applicability of the model. For example, a true binding region in a biological experiment for protein-DNA-binding depends on the properties of the protein, such as flexibility, binding-strength, molecular conformation and/or the interaction with other molecules. The signal and size of the binding regions are therefore affected by the properties of the protein. Thus the creation of in silico data set requires many well characterized (validated) binding regions among different types of biological experiments to model true-binding regions.

Coming back to the initial question, "*What is the best method to use?*" appears to be a too generic question as the choice of the best method depends on the exact research question. As an example, if one is interested in the genes that are in close proximity of the top enriched candidate binding regions, the majority of methodologies will likely give very similar results. However differences in results between methodologies may occur if one is interested in the motifs (instead of neighboring genes) among the (top) detected binding regions.

Motif analysis is a straight-forward approach to benchmark the detected binding regions determined using a particular methodology. It could for instance address the question whether expected consensus binding sites are detected. Such findings may help deciding which methodology to choose. It is important to note that in the end, laboratory experiments are indispensable to demonstrate the biological significance of particular detected region. For instance in case transcription factor binding regions are identified by means ChIP-on-chip or ChIP-Seq, these interactions should be validated using ChIP-PCR and possibly using a combination in functional experiments to demonstrate the real meaning of such interactions.

The advantage of our methodology is that it detects candidate regions without defining a priori the expected size of the ROI, and is therefore applicable for different biological experiments (protein-DNA-binding, DNA-methylation, histone modifications). Although we are able to easily follow-up biological relevance of the detected binding regions, data integration (e.g. with gene expression profiles) may be the key to gain more insights in the disease state as these approaches can speed up the identification of critical leukemogenic hits.

IDENTIFICATION OF C/EBPα TARGETS BY PROMOTER BINDING AND mRNA CHANGES

A powerful technology to detect potential functional regions on the genome is by using tiling-array (e.g. ChIP-on-chip) or by massive parallel DNA sequencing (e.g. ChIP-seq) data. The binding of proteins to the DNA does not necessarily result in changes of mRNA expression levels (relative gene activation or repression can be measured with the use of, e.g. gene expression profiling). Integration of e.g. ChIP-on-chip or ChIP-seq with gene expression profiles (GEP) is of great importance to get better understanding in the involved pathways and the exact functional role of transcription factors on gene expression regulation (such as for *CEBPA*). Such results can then be instrumental for the development

of novel targeted therapies for specific subtypes in AML. Although ChIP-on-chip or ChIP-seq and gene expression data sets are widely used, it is not straightforward to integrate data sets as these procedures are not standard.

In Chapter 6 we provide data to demonstrate that C/EBPα may not only act as an activator but also as a repressor of transcription. We proposed that absence of C/EBPα is one of the transforming events towards mixed myeloid/T-lymphoid leukemia. In particular T-cell related genes, that are upregulated when C/EBPα is absent, are predicted to be targets of C/EBPα and repressed in the presence of this transcription factor. Such transformation effect is likely to be caused by a very complex network of relationships between proteins, motifs and epigenetic regulation, such as DNA-methylation. In addition, our results are detected in a 32D cell line model, and may not be identical to the human situation. The use of (public) ChIP-on-chip or ChIP-Seq data sets of the identified binding-proteins, gene expression and DNA-methylation profiles in large cohorts of AML samples will undoubtedly contribute to further insights.

THE IMPORTANCE OF INTEGRATIVE AND COMBINATORIAL ANALYSES OF GEP AND DMP

The use of gene expression profiling (GEP), to measure changes in mRNA expression levels in patient samples, has become very important in cancer research because of its wide applicability[18,24,61,62]. An example is the discovery of previously unrecognized subgroups[18]. However, measurements of mRNA expression levels is only one type of disturbance that may occur in malignant cells. Another type of disturbance in malignant cells are changes in DNA-methylation[249], which can be measured in patient samples using DNA-methylation profiling (DMP)[250]. Relations between DNA-methylation and mRNA expression in AML are known to exist[251], however no comprehensive integrative analysis has been performed between GEP and DMP so far in AML. Combining these data sets may result in unique cancerous patterns that characterizes subgroups in AML. There are two major challenges: the development of a statistical approach to combine both data sets, and secondly the identification of novel subgroups in AML which are not artifacts induced by the statistical approach. In Chapter 9 we show with an unsupervised analysis that the combined data sets revealed two unrecognized subgroups in AML. These two subgroups could not be identified using GEP or DMP alone[18,58].

There are still questions to be addressed to learn more about the biological relations between gene expression and DNA-methylation. Question such as, "*how does DNA-methylation influences the change of gene expression for a specific AML subtype?*" will undoubtedly contribute to further insights. This is particularly of interest for patient groups that have malignant cells affected by both abnormal gene expression and DNA-methylation levels.

EVALUATION OF TREATMENT EFFECTS FOR PATIENT SUBGROUPS REQUIRES LARGE STUDY-SIZES

The identification of different subtypes in AML is especially of importance as it can be used in treatment protocols[35,228,252,253]. We needed a large cohort of 1182 normal karyotype patients to show that favourable outcome is mainly observed in AML with *CEBPA*^{dm} and not in *CEBPA*^{sm} [178] (Chapter 7). In addition, we showed that concurrent mutations occur more frequently in *CEBPA*^{sm} than in *CEBPA*^{dm} AML and affects outcome, e.g. *CEBPA*^{sm} is dominated by concurrent *NPM1* and/or *FLT3* internal tandem duplication (ITD) mutations. We therefore proposed that only *CEBPA*^{dm} should be considered as a separate entity in the classification of AMLs. The importance of *CEBPA*^{dm} as a separate entity is stressed by the notion that various genetic and cytogenetic abnormalities in AML have clinical value as they may predict disease outcome and response to treatment[254]. To address this question, thus the impact of allogeneic and autologous Hematopoietic Stem Cell Transplantation (alloHSCT, autoHSCT respectively) in comparison to chemotherapy consolidation, required even more patient samples. By combining clinical data from Dutch-Belgian-Swiss Hemato-Oncology Cooperative Group (HOVON/SAKK) and German-Austrian AML Study Group (AMLSG) trials we compiled 5724 patients. The number of patients that accomplished to our selection was however only 124 *CEBPA*^{dm} patients (Chapter 8). These were subsequently split into smaller subgroups based on how these patients were treated. A challenge in the analysis is to overcome the time-to-treatment bias from the no-transplant group to the alloHSCT or autoHSCT groups at the time-point of HSCT. Without correction for the time-to-treatment bias, the results in terms of outcome may be misleading or even incorrect because patients that received alloHSCT or autoHSCT treatment are consolidated after chemotherapy. We were able to show that patients receiving an alloHSCT or autoHSCT in first CR, showed significantly less relapses compared to patients receiving chemotherapy. Meaning that relapse-free survival (RFS) was significantly superior in patients receiving an alloHSCT or autoHSCT in first CR compared to patients receiving chemotherapy. However, the superior RFS after alloHSCT and autoHSCT did not translate in a significant better overall survival (OS). This may be caused due to a high second complete remission rate for patients that received intensive chemo treatment. Thirty-eight out of 70 patients relapsed after intensive chemo therapy, and only 6 out of 38 patients relapsed after receiving alloHSCT or autoHSCT treatment. It seems that alloHSCT is a good option in first CR, but an alternative and not unreasonable strategy would be to postpone the alloHSCT in first CR and keep the option of alloHSCT for salvage for the restricted fraction of patients after relapse. Our data supports both strategies, on the one hand an alloHSCT or autoHSCT in first CR and on the other hand intensive chemotherapy as consolidation in first CR and an alloHSCT in case of relapse. An important notion is however that patients have to be well informed about the risks and consequences of alloHSCT and autoHSCT in first CR. Recent findings showed frequently associated of *CEBPA*^{dm} with *GATA2* zinc finger 1 mutations[255,256], and that *GATA2* mutations are associated with favorable outcome compared to other *GATA2*$^{wild\text{-}type}$ genotypes. Although no significant differences in outcome were observed between the groups *GATA2*^{mutant}/*CEBPA*^{dm} versus *GATA2*$^{wild\text{-}type}$/*CEBPA*^{dm}, it requires further investigation with even larger cohorts to investigate the exact role of mutant *GATA2* on top of *CEBPA*^{dm} in terms of patients receiving an alloHSCT or autoHSCT in first CR.

NEXT-GENERATION SEQUENCING AND THE DISCOVERY OF AML SUBTYPES

The ultimate goal in cancer research is the development of personalized treatment based on the specific cancerous patterns in a single patient. The introduction of next-generation sequencing technology[257] may be a first step in this process as it allows us to accurately characterize all the abnormalities for a single patient. However, the generated data that need to be processed, analysed and confirmed may remain a challenge for the coming decennia. So far, microarray experiments (with relatively limited resolutions compared to NGS data) are intensively used to find groups of patients with similar cancerous patterns (denoted as subtype) [18]. Although many subtypes are already identified and recognized by the World Health Organization[228], it is highly conceivable that more AML subtypes do exist but are hidden due to the complex relationships of molecules in cancerous cells.

A challenge is to find these AML subtypes and to infer whether these subgroups are "true" and not artefacts induced by the inherent tendency of the probabilistic models. A "true" subtype of AML is defined as a group of samples that can be identified with common features, such as common molecular aberrations and/or clinical responses. Ideally, AML subtypes can be discovered by: 1. analyzing various data sources of the same samples in an integrative manner, such as gene expression, DNA-methylation, microRNAs and pathways. 2. Using high data-resolutions, and 3. Using a high number of patient data.

The use of NGS technology will cover the first two points. The third point is crucial as it may reveal the existence of subtypes that are only seen in low frequencies, and importantly it will increase the computational power to find with high confidence (novel) AML subtypes. So far, small numbers of patient samples is the most important limiting factor in integrative studies, and for the detection of AML subtypes. As an example, if multiple data sets are combined (point 1) that have high genomic coverage (point 2), it will result in an exploding number of features that heavily increases the number of computations (e.g. degrees of freedom increases). This suggests that it becomes harder to find true positive relationships if only point 1 and 2 are followed. This can particularly become a problem as we enter the sequencing era where the genome coverage is massively increased without massively increasing the number of samples.

A key to continue expanding our knowledge about diseases such as leukemia (with the use of NGS data) is to share data across the world and to make IT frameworks accessible for data sharing and computations. Without such as an approach, and under the assumption of continuous increase of high data-resolutions but steady (low) sample sizes, the balance is disturbed and leads to the "*curse of dimensionality*[258]". Which may result in a dramatically increase of false positives. Validation of the results will be stressed even more.

NGS technology can nowadays be of great value in many other applications such as for the detection of gene mutation or the determination of binding regions of proteins, but also to support the analysis of conventional experiments (tiling-arrays) in applications such as the detection of AML subtypes (Chapter 9). In contradiction to the statistical approaches that are known for tiling-array experiments, NGS data analysis has started very recently. The required

statistical approaches are still immature and need to be developed, optimized, and validated. A new statistical-era for NGS data analysis has just been started.

REFERENCES

1. Bruce Alberts AJ, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. Molecular Biology of the Cell.Fourth Edition.
2. Vickaryous MK, Hall BK. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc.* Aug 2006;81(3):425-455.
3. Pennisi E. Genomics. DNA study forces rethink of what it means to be a gene. *Science.* Jun 15 2007;316(5831):1556-1557.
4. Pearson H. Genetics: what is a gene? *Nature.* May 25 2006;441(7092):398-401.
5. Crick F. Central dogma of molecular biology. *Nature.* Aug 8 1970;227(5258):561-563.
6. Finishing the euchromatic sequence of the human genome. *Nature.* Oct 21 2004;431(7011):931-945.
7. Butte AJ, Dzau VJ, Glueck SB. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". *Physiol Genomics.* Dec 21 2001;7(2):95-96.
8. Eisenberg E, Levanon EY. Human housekeeping genes are compact. *Trends Genet.* Jul 2003;19(7):362-365.
9. Howlader N NA, Krapcho M, Neyman N, Aminou R, Waldron W, Altekruse SF, Kosary CL, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Chen HS, Feuer EJ, Cronin KA, Edwards BK (eds). SEER Cancer Statistics Review, 1975-2008, National Cancer Institute. http://seer.cancer.gov/csr/1975_2008/. 2011.
10. Klein CA. Cancer. The metastasis cascade. *Science.* Sep 26 2008;321(5897):1785-1787.
11. Evan G, Littlewood T. A matter of life and cell death. *Science.* Aug 28 1998;281(5381):1317-1322.
12. Weinberg RA. how cancer arises. *Sci. Am.* 1996;275(3):62-70.
13. Croce CM. Oncogenes and cancer. *N Engl J Med.* Jan 31 2008;358(5):502-511.
14. Sherr CJ. Principles of tumor suppression. *Cell.* Jan 23 2004;116(2):235-246.
15. Simpson AJ. The natural somatic mutation frequency and human carcinogenesis. *Adv Cancer Res.* 1997;71:209-240.
16. Muller-Sieburg CE, Cho RH, Thoman M, Adkins B, Sieburg HB. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood.* Aug 15 2002;100(4):1302-1309.
17. McCulloch EA. Stem cell renewal and determination during clonal expansion in normal and leukaemic haemopoiesis. *Cell Prolif.* Sep 1993;26(5):399-425.
18. 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.
19. Kottaridis PD, Gale RE, Frew ME, et al. The presence of a FLT3 internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy: analysis of 854 patients from the United Kingdom Medical Research Council AML 10 and 12 trials. *Blood.* Sep 15 2001;98(6):1752-1759.
20. Kottaridis PD, Gale RE, Linch DC. Prognostic implications of the presence of FLT3 mutations in patients with acute myeloid leukemia. *Leuk Lymphoma.* Jun 2003;44(6):905-913.
21. Thiede C, Steudel C, Mohr B, et al. Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. *Blood.* Jun 15 2002;99(12):4326-4335.
22. Frohling S, Schlenk RF, Breitruck J, et al. Prognostic significance of activating FLT3 mutations in younger adults (16 to 60 years) with acute myeloid leukemia and normal cytogenetics: a study of the AML Study Group Ulm. *Blood.* Dec 15 2002;100(13):4372-4380.
23. Whitman SP, Archer KJ, Feng L, et al. Absence of the wild-type allele predicts poor prognosis in adult de novo acute myeloid leukemia with normal cytogenetics and the internal tandem duplication of FLT3: a cancer and leukemia group B study. *Cancer Res.* Oct 1 2001;61(19):7233-7239.
24. 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.
25. Falini B, Nicoletti I, Bolli N, et al. Translocations and mutations involving the nucleophosmin (NPM1) gene in lymphomas and leukemias. *Haematologica.* Apr 2007;92(4):519-532.
26. Lekstrom-Himes J, Xanthopoulos KG. Biological role of the CCAAT/enhancer-binding protein family of transcription factors. *J Biol Chem.* Oct 30 1998;273(44):28545-28548.

27. 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.
28. 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.
29. Harris NL, Jaffe ES, Diebold J, et al. World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee meeting-Airlie House, Virginia, November 1997. *J Clin Oncol.* Dec 1999;17(12):3835-3849.
30. Lowenberg B. Diagnosis and prognosis in acute myeloid leukemia--the art of distinction. *N Engl J Med.* May 1 2008;358(18):1960-1962.
31. 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.
32. Bennett JM, Catovsky D, Daniel MT, et al. Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British Cooperative Group. *Ann Intern Med.* Oct 1985;103(4):620-625.
33. Lowenberg B, Downing JR, Burnett A. Acute myeloid leukemia. *N Engl J Med.* Sep 30 1999;341(14):1051-1062.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. Mueller BU, Pabst T. C/EBPalpha and the pathophysiology of acute myeloid leukemia. *Curr Opin Hematol.* Jan 2006;13(1):7-14.
40. Nerlov C. C/EBPalpha mutations in acute myeloid leukaemias. *Nat Rev Cancer.* May 2004;4(5):394-400.
41. 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.
42. 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.
43. 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.
44. 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.* Feb 24 2011;117(8):2469-2475.
45. 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.
46. 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.

47. 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.
48. 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.
49. Anastas JN, Moon RT. WNT signalling pathways as therapeutic targets in cancer. *Nat Rev Cancer.* Jan 2013;13(1):11-26.
50. 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.
51. van Vliet MH, Burgmer P, de Quartel L, et al. Detection of CEBPA Double-mutants in Acute Myeloid Leukemia Using a Custom Gene Expression Array. *GeneT-test Mol Biomarkers.* Mar 13 2013.
52. Ramji DP, Foka P. CCAAT/enhancer-binding proteins: structure, function and regulation. *Biochem J.* Aug 1 2002;365(Pt 3):561-575.
53. Ossipow V, Descombes P, Schibler U. CCAAT/enhancer-binding protein mRNA is translated into multiple proteins with different transcription activation potentials. *Proc Natl Acad Sci U S A.* Sep 1 1993;90(17):8219-8223.
54. Wolfler A, Danen-van Oorschot AA, Haanstra JR, et al. Lineage-instructive function of C/EBPalpha in multipotent hematopoietic cells and early thymic progenitors. *Blood.* Nov 18 2010;116(20):4116-4125.
55. Otto C, Reiche K, Hackermuller J. Detection of differentially expressed segments in tiling-array data. *Bioinformatics.* Jun 1 2012;28(11):1471-1479.
56. Beekman R, Valkhof M, Erkeland SJ, et al. Retroviral integration mutagenesis in mice and comparative analysis in human AML identify reduced PTP4A3 expression as a prognostic indicator. *PLoS One.* 2011;6(10):e26537.
57. Kim CK, Kikuchi S, Hahn JH, Park SC, Kim YH, Lee BW. Computational identification of anthocyanin-specific transcription factors using a rice microarray and maximum boundary range algorithm. *Evol Bioinform Online.* 2010;6:133-141.
58. Figueroa ME, Lugthart S, Li Y, et al. DNA-methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell.* Jan 19 2010;17(1):13-27.
59. Mueller BU, Pabst T. Lineage-specific transcription factor aberrations in AML. *Cancer Treat Res.* 2010;145:109-125.
60. Meyers S, Downing JR, Hiebert SW. Identification of AML-1 and the (8;21) translocation protein (AML-1/ETO) as sequence-specific DNA-binding proteins: the runt homology domain is required for DNA-binding and protein-protein interactions. *Mol Cell Biol.* Oct 1993;13(10):6336-6345.
61. 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.* Nov 15 2007;110(10):3706-3714.
62. Verhaak RG, Wouters BJ, Erpelinck CA, et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica.* Jan 2009;94(1):131-134.
63. What You Need To Know About™ Leukemia. 2008.
64. Aparicio O, Geisberg J, Struhl K. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Cell Biol.* 2004;Chapter 17:Unit 17.17.
65. Liu X. Getting started in tiling microarray analysis. *PLoS Comput Biol.* 2007;3(10):1842 - 1844.
66. Weber M, Davies J, Wittig D, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA-methylation in normal and transformed human cells. *Nat Genet.* 2005;37(8):853 - 862.
67. Bertone P, Stolc V, Royce TE, et al. Global identification of human transcribed sequences with genome tiling-arrays. *Science.* Dec 24 2004;306(5705):2242-2246.
68. Crawford G, Davis S, Scacheri P, et al. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature methods.* 2006;3(7):503 - 509.
69. Heidenblad M, Lindgren D, Jonson T, et al. Tiling resolution array CGH and high density expression profiling of urothelial carcinomas delineate genomic amplicons and candidate target genes specific for advanced tumors. *BMC Med Genomics.* 2008;1:3.

70. Royce T, Rozowsky J, Bertone P, et al. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.* 2005;21(8):466 - 475.
71. Keles S, Laan M, Dudoit S, Cawley S. Multiple testing methods for ChIP-on-chip high density oligonucleotide array data. *J Comput Biol.* 2006;13(3):579 - 613.
72. Li W, Meyer C, Liu X. A hidden Markov model for analyzing ChIP-on-chip experiments on genome tiling-arrays and its application to p53 binding sequences. *Bioinformatics.* 2005;21(Suppl 1):i274 - i282.
73. Ji H, Wong W. TileMap: create chromosomal map of tiling-array hybridizations. *Bioinformatics.* 2005;21(18):3629 - 3636.
74. Johnson WE, Li W, Meyer CA, et al. Model-based analysis of tiling-arrays for ChIP-on-chip. *Proc Natl Acad Sci U S A.* Aug 15 2006;103(33):12457-12462.
75. Sun W, Buck M, Patel M, Davis I. Improved ChIP-on-chip analysis by a mixture model approach. *BMC Bioinformatics.* 2009;10:173.
76. Kuan P, Chun H, Keles S. CMARRT: a tool for the analysis of ChIP-on-chip data from tiling-arrays by incorporating the correlation structure. *Pac Symp Biocomput.* 2008:515 - 526.
77. Zacher B, Kuan P, Tresch A. Starr: Simple Tiling-array analysis of Affymetrix ChIP-on-chip data. *BMC Bioinformatics.* 2010;11:194.
78. Toedling J, Skylar O, Sklyar O, et al. Ringo-an R/Bioconductor package for analyzing ChIP-on-chip readouts. *BMC Bioinformatics.* 2007;8:221.
79. Ji X, Li W, Song J, Wei L, Liu XS. CEAS: cis-regulatory element annotation system. *Nucleic Acids Res.* Jul 1 2006;34(Web Server issue):W551-554.
80. Tinel M, Berson A, Elkahwaji J, Cresteil T, Beaune P, Pessayre D. Downregulation of cytochromes P450 in growth-stimulated rat hepatocytes: role of c-Myc induction and impaired C/EBP binding to DNA. *J Hepatol.* 2003;39(2):171 - 178.
81. Wang W, Wang X, Ward A, Touw I, Friedman A. C/EBPalpha and G-CSF receptor signals cooperate to induce the myeloperoxidase and neutrophil elastase genes. *Leukemia.* 2001;15(5):779 - 786.
82. Zhang P, Iwama A, Datta MW, Darlington GJ, Link DC, Tenen DG. Upregulation of interleukin 6 and granulocyte colony-stimulating factor receptors by transcription factor CCAAT enhancer binding protein alpha (C/EBP alpha) is critical for granulopoiesis. *J Exp Med.* Sep 21 1998;188(6):1173-1184.
83. Erkeland S, Valkhof M, Heijmans-Antonissen C, et al. Large scale identification of disease genes involved in acute myeloid leukemia. *J Virol.* 2004;78(4):1971 - 1980.
84. Touw I, Erkeland S. Retroviral insertion mutagenesis in mice as a comparative oncogenomics tool to identify disease genes in human leukemia. *Mol Ther.* 2007;15:13 - 19.
85. Theodorou V, Kimm M, Boer M, et al. MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer. *Nat Genet.* 2007;39(6):759 - 769.
86. Suzuki T, Shen H, Akagi K, et al. New genes involved in cancer identified by retroviral tagging. *Nat Genet.* 2002;32:166 - 174.
87. Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR. Applications of DNA tiling-arrays for whole-genome analysis. *Genomics.* Jan 2005;85(1):1-15.
88. 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.* Jan 22 2003;19(2):185-193.
89. Taskesen E, Beekman R, de Ridder J, et al. HAT: hypergeometric analysis of tiling-arrays with application to promoter-GeneChip data. *BMC Bioinformatics.* 2010;11:275.
90. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.
91. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics.* Aug 1 2008;24(15):1729-1730.
92. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-on-chip and ChIP-seq data. *Nature biotechnology.* Nov 2008;26(11):1293-1300.
93. Valouev A, Johnson DS, Sundquist A, et al. Genome wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods.* Sep 2008;5(9):829-834.
94. Rozowsky J, Euskirchen G, Auerbach RK, et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology.* Jan 2009;27(1):66-75.

95. Wilbanks E.G., Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One.* 2010;5(7):e11471.
96. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* Sep 2009;19(9):1639-1645.
97. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome wide expression profiles. *Proc Natl Acad Sci U S A.* Oct 25 2005;102(43):15545-15550.
98. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res.* Nov 1 2001;125(1-2):279-284.
99. Yosef H. A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika.* 1988;75(4):2.
100. N Cristianini MWH. Introduction to Computational Genomics - A Case Studies Approach. *Cambridge University Press.* 2007.
101. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* Jul 15 2009;25(14):1754-1760.
102. Bonora-Centelles A, Jover R, Mirabet V, et al. Sequential hepatogenic transdifferentiation of adipose tissue-derived stem cells: relevance of different extracellular signaling molecules, transcription factors involved, and expression of new key marker genes. *Cell transplantation.* 2009;18(12):1319-1340.
103. Hon GC, Hawkins RD, Ren B. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet.* Oct 15 2009;18(R2):R195-201.
104. Wu CY, Tsai YP, Wu MZ, Teng SC, Wu KJ. Epigenetic reprogramming and post-transcriptional regulation during the epithelial-mesenchymal transition. *Trends Genet.* Sep 2012;28(9):454-463.
105. Hartigan JAHaPM. The Dip Test of Unimodality. *The Annals of Statistics.* 1985;13(1):14.
106. Sati S, Tanwar VS, Kumar KA, et al. High resolution methylome map of rat indicates role of intragenic DNA-methylation in identification of coding region. *PLoS One.* 2012;7(2):e31621.
107. Stitzel ML, Sethupathy P, Pearson DS, et al. Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* Nov 3 2010;12(5):443-455.
108. Li N, Ye M, Li Y, et al. Whole genome DNA-methylation analysis based on high throughput sequencing technology. *Methods.* Nov 2010;52(3):203-212.
109. Robertson G, Hirst M, Bainbridge M, et al. Genome wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods.* Aug 2007;4(8):651-657.
110. Ehret GB, Reichenbach P, Schindler U, et al. DNA-binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites. *J Biol Chem.* Mar 2 2001;276(9):6675-6688.
111. Gunaje JJ, Bhat GJ. Involvement of tyrosine phosphatase PTP1D in the inhibition of interleukin-6-induced Stat3 signaling by alpha-thrombin. *Biochem Biophys Res Commun.* Oct 19 2001;288(1):252-257.
112. 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.
113. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. *Genome Biol.* 2007;8(2):R24.
114. Li X, Leung S, Qureshi S, Darnell JE, Jr., Stark GR. Formation of STAT1-STAT2 heterodimers and their role in the activation of IRF-1 gene transcription by interferon-alpha. *J Biol Chem.* Mar 8 1996;271(10):5790-5794.
115. Chatterjee-Kishore M, van Den Akker F, Stark GR. Adenovirus E1A downregulates LMP2 transcription by interfering with the binding of stat1 to IRF1. *J Biol Chem.* Jul 7 2000;275(27):20406-20411.
116. Takeda A, Hamano S, Yamanaka A, et al. Cutting edge: role of IL-27/WSX-1 signaling for induction of T-bet through activation of STAT1 during initial Th1 commitment. *J Immunol.* May 15 2003;170(10):4886-4890.
117. Xie B, Zhao J, Kitagawa M, et al. Focal adhesion kinase activates Stat1 in integrin-mediated cell migration and adhesion. *J Biol Chem.* Jun 1 2001;276(22):19512-19523.
118. Li X, Leung S, Kerr IM, Stark GR. Functional subdomains of STAT2 required for preassociation with the alpha interferon receptor and for signaling. *Mol Cell Biol.* Apr 1997;17(4):2048-2056.
119. Swerdlow S, Campo E, Harris N, et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. *Lyon, IARC Press* 2008.
120. Lowenberg B. Acute myeloid leukemia: the challenge of capturing disease variety. *Hematology Am Soc Hematol Educ Program.* 2008:1-11.
121. Wouters BJ, Lowenberg B, Delwel R. A decade of genome wide gene expression profiling in acute myeloid leukemia: flashback and prospects. *Blood.* Jan 8 2009;113(2):291-298.

122. Kool J, Berns A. High throughput insertional mutagenesis screens in mice to identify oncogenic networks. *Nat Rev Cancer.* Jun 2009;9(6):389-399.
123. Suzuki T, Minehata K, Akagi K, Jenkins NA, Copeland NG. Tumor suppressor gene identification using retroviral insertional mutagenesis in Blm-deficient mice. *EMBO J.* Jul 26 2006;25(14):3422-3431.
124. Lorincz MC, Schubeler D, Goeke SC, Walters M, Groudine M, Martin DI. Dynamic analysis of proviral induction and De Novo methylation: implications for a histone deacetylase-independent, methylation density-dependent mechanism of transcriptional repression. *Mol Cell Biol.* Feb 2000;20(3):842-850.
125. Yao S, Sukonnik T, Kean T, Bharadwaj RR, Pasceri P, Ellis J. Retrovirus silencing, variegation, extinction, and memory are controlled by a dynamic interplay of multiple epigenetic modifications. *Mol Ther.* Jul 2004;10(1):27-36.
126. Swindle CS, Kim HG, Klug CA. Mutation of CpGs in the murine stem cell virus retroviral vector long terminal repeat represses silencing in embryonic stem cells. *J Biol Chem.* Jan 2 2004;279(1):34-41.
127. Voisin V, Barat C, Hoang T, Rassart E. Novel insights into the pathogenesis of the Graffi murine leukemia retrovirus. *J Virol.* Apr 2006;80(8):4026-4037.
128. Weber M, Davies JJ, Wittig D, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA-methylation in normal and transformed human cells. *Nat Genet.* Aug 2005;37(8):853-862.
129. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. *Science.* Jun 13 2003;300(5626):1749-1751.
130. Basak S, Jacobs SB, Krieg AJ, et al. The metastasis-associated gene Prl-3 is a p53 target involved in cell-cycle regulation. *Mol Cell.* May 9 2008;30(3):303-314.
131. Min SH, Kim DM, Heo YS, Kim HM, Kim IC, Yoo OJ. Downregulation of p53 by phosphatase of regenerating liver 3 is mediated by MDM2 and PIRH2. *Life sciences.* Jan 2 2010;86(1-2):66-72.
132. Pfeifer D, Wallin A, Holmlund B, Sun XF. Protein expression following gamma-irradiation relevant to growth arrest and apoptosis in colon cancer cells. *Journal of cancer research and clinical oncology.* Nov 2009;135(11):1583-1592.
133. Hao RT, Zhang XH, Pan YF, et al. Prognostic and metastatic value of phosphatase of regenerating liver-3 in invasive breast cancer. *Journal of cancer research and clinical oncology.* Sep 2010;136(9):1349-1357.
134. Laurent C, Valet F, Planque N, et al. High PTP4A3 phosphatase expression correlates with metastatic risk in uveal melanoma patients. *Cancer Res.* Feb 1 2011;71(3):666-674.
135. Mollevi DG, Aytes A, Padulles L, et al. PRL-3 is essentially overexpressed in primary colorectal tumours and associates with tumour aggressiveness. *Br J Cancer.* Nov 18 2008;99(10):1718-1725.
136. Wang Z, Cai SR, He YL, et al. Elevated PRL-3 expression was more frequently detected in the large primary gastric cancer and exhibits a poor prognostic impact on the patients. *Journal of cancer research and clinical oncology.* Aug 2009;135(8):1041-1046.
137. Zhao WB, Li Y, Liu X, Zhang LY, Wang X. Evaluation of PRL-3 expression, and its correlation with angiogenesis and invasion in hepatocellular carcinoma. *International journal of molecular medicine.* Aug 2008;22(2):187-192.
138. Wang H, Quah SY, Dong JM, Manser E, Tang JP, Zeng Q. PRL-3 downregulates PTEN expression and signals through PI3K to promote epithelial-mesenchymal transition. *Cancer Res.* Apr 1 2007;67(7):2922-2926.
139. Juric D, Lacayo NJ, Ramsey MC, et al. Differential gene expression patterns and interaction networks in BCR-ABL-positive and -negative adult acute lymphoblastic leukemias. *J Clin Oncol.* Apr 10 2007;25(11):1341-1349.
140. Broyl A, Hose D, Lokhorst H, et al. Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. *Blood.* Oct 7 2010;116(14):2543-2553.
141. Fagerli UM, Holt RU, Holien T, et al. Overexpression and involvement in migration by the metastasis-associated phosphatase PRL-3 in human myeloma cells. *Blood.* Jan 15 2008;111(2):806-815.
142. Zhou J, Bi C, Chng WJ, et al. PRL-3, a Metastasis Associated Tyrosine Phosphatase, Is Involved in FLT3-ITD Signaling and Implicated in Anti-AML Therapy. *PLoS One.* 2011;6(5):e19798.
143. Good SR, Thieu VT, Mathur AN, et al. Temporal induction pattern of STAT4 target genes defines potential for Th1 lineage-specific programming. *J Immunol.* Sep 15 2009;183(6):3839-3847.
144. Crawford GE, Davis S, Scacheri PC, et al. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods.* Jul 2006;3(7):503-509.
145. Yamada K, Lim J, Dale JM, et al. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science.* Oct 31 2003;302(5646):842-846.

146. Munch K, Gardner PP, Arctander P, Krogh A. A hidden Markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics.* 2006;7:239.
147. Yu WH, Hovik H, Chen T. A hidden Markov support vector machine framework incorporating profile geometry learning for identifying microbial RNA in tiling-array data. *Bioinformatics.* Jun 1 2010;26(11):1423-1430.
148. Knott SR, Viggiani CJ, Aparicio OM, Tavare S. Strategies for analyzing highly enriched IP-chip datasets. *BMC Bioinformatics.* 2009;10:305.
149. Kuan PF, Chun H, Keles S. CMARRT: a tool for the analysis of ChIP-on-chip data from tiling-arrays by incorporating the correlation structure. *Pac Symp Biocomput.* 2008:515-526.
150. Zhang Y. Poisson approximation for significance in genome wide ChIP-on-chip tiling-arrays. *Bioinformatics.* Dec 15 2008;24(24):2825-2831.
151. Wu H, Ji H. JAMIE: A software tool for jointly analyzing multiple ChIP-on-chip experiments. *Methods Mol Biol.* 2012;802:363-375.
152. Mo Q, Liang F. A hidden Ising model for ChIP-on-chip data analysis. *Bioinformatics.* Mar 15 2010;26(6):777-783.
153. Mo Q, Liang F. Bayesian modeling of ChIP-on-chip data through a high-order Ising model. *Biometrics.* Dec 2010;66(4):1284-1294.
154. Karpikov A, Rozowsky J, Gerstein M. Tiling-array data analysis: a multiscale approach using wavelets. *BMC Bioinformatics.* 2011;12:57.
155. Lan X, Bonneville R, Apostolos J, Wu W, Jin VX. W-ChIPeaks: a comprehensive web application tool for processing ChIP-on-chip and ChIP-seq data. *Bioinformatics.* Feb 1 2011;27(3):428-430.
156. Kechris KJ, Biehs B, Kornberg TB. Generalizing moving averages for tiling-arrays using combined P-value statistics. *Stat Appl Genet Mol Biol.* 2010;9(1):Article29.
157. Zacher B, Kuan PF, Tresch A. Starr: Simple Tiling-array analysis of Affymetrix ChIP-on-chip data. *BMC Bioinformatics.* 2010;11:194.
158. Droit A, Cheung C, Gottardo R. rMAT--an R/Bioconductor package for analyzing ChIP-on-chip experiments. *Bioinformatics.* Mar 1 2010;26(5):678-679.
159. Judy JT, Ji H. TileProbe: modeling tiling-array probe effects using publicly available data. *Bioinformatics.* Sep 15 2009;25(18):2369-2375.
160. Cesaroni M, Cittaro D, Brozzi A, Pelicci PG, Luzi L. CARPET: a web-based package for the analysis of ChIP-on-chip and expression tiling data. *Bioinformatics.* Dec 15 2008;24(24):2918-2920.
161. Lieberman LA, Banica M, Reiner SL, Hunter CA. STAT1 plays a critical role in the regulation of antimicrobial effector mechanisms, but not in the development of Th1-type responses during toxoplasmosis. *J Immunol.* Jan 1 2004;172(1):457-463.
162. Nguyen KB, Watford WT, Salomon R, et al. Critical role for STAT4 activation by type 1 interferons in the interferon-gamma response to viral infection. *Science.* Sep 20 2002;297(5589):2063-2066.
163. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* Jul 1 2003;31(13):3576-3579.
164. Kel A, Voss N, Jauregui R, Kel-Margoulis O, Wingender E. Beyond microarrays: find key transcription factors controlling signal transduction pathways. *BMC Bioinformatics.* 2006;7 Suppl 2:S13.
165. Rosenbauer F, Tenen DG. Transcription factors in myeloid development: balancing differentiation with transformation. *Nat Rev Immunol.* Feb 2007;7(2):105-117.
166. Friedman AD. Transcriptional control of granulocyte and monocyte development. *Oncogene.* Oct 15 2007;26(47):6816-6828.
167. de Bruijn MF, Speck NA. Core-binding factors in hematopoiesis and immune function. *Oncogene.* May 24 2004;23(24):4238-4248.
168. Rabault B, Ghysdael J. Calcium-induced phosphorylation of ETS1 inhibits its specific DNA-binding activity. *J Biol Chem.* Nov 11 1994;269(45):28143-28151.
169. Tenen DG, Hromas R, Licht JD, Zhang DE. Transcription factors, normal myeloid development, and leukemia. *Blood.* Jul 15 1997;90(2):489-519.
170. Grall FT, Prall WC, Wei W, et al. The Ets transcription factor ESE-1 mediates induction of the COX-2 gene by LPS in monocytes. *FEBS J.* Apr 2005;272(7):1676-1687.
171. Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC. Getting started in biological pathway construction and analysis. *PLoS Comput Biol.* Feb 2008;4(2):e16.

172. Descombes P, Schibler U. A liver-enriched transcriptional activator protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA. *Cell.* Nov 1 1991;67(3):569-579.
173. Lin FT, MacDougald OA, Diehl AM, Lane MD. A 30-kDa alternative translation product of the CCAAT/enhancer binding protein alpha message: transcriptional activator lacking antimitotic activity. *Proc Natl Acad Sci U S A.* Oct 15 1993;90(20):9606-9610.
174. Welm AL, Timchenko NA, Darlington GJ. C/EBPalpha regulates generation of C/EBPbeta isoforms through activation of specific proteolytic cleavage. *Mol Cell Biol.* Mar 1999;19(3):1695-1704.
175. 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.
176. Pabst T, Mueller BU, Harakawa N, et al. AML1-ETO downregulates the granulocytic differentiation factor C/EBPalpha in t(8;21) myeloid leukemia. *Nature medicine.* Apr 2001;7(4):444-451.
177. Wolfler A, Danen-van Oorschot AA, Haanstra JR, et al. Lineage-instructive function of C/EBPalpha in multipotent hematopoietic cells and early thymic progenitors. *Blood.* Nov 18;116(20):4116-4125.
178. 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: further evidence for CEBPA double-mutant AML as a distinctive disease entity. *Blood.* Dec 21.
179. 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.* Jun 1;28(16):2739-2747.
180. 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.
181. 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.
182. Taskesen E, Beekman R, de Ridder J, et al. HAT: hypergeometric analysis of tiling-arrays with application to promoter-GeneChip data. *BMC Bioinformatics.*11:275.
183. 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.* Dec 1993;13(12):7774-7781.
184. Alberich-Jorda M, Wouters B, Balastik M, et al. C/EBPgamma deregulation results in differentiation arrest in acute myeloid leukemia. *J Clin Invest.* Dec 3 2012;122(12):4490-4504.
185. Smith LT, Hohaus S, Gonzalez DA, Dziennis SE, Tenen DG. PU.1 (Spi-1) and C/EBP alpha regulate the granulocyte colony-stimulating factor receptor promoter in myeloid cells. *Blood.* Aug 15 1996;88(4):1234-1247.
186. Hohaus S, Petrovick MS, Voso MT, Sun Z, Zhang DE, Tenen DG. PU.1 (Spi-1) and C/EBP alpha regulate expression of the granulocyte-macrophage colony-stimulating factor receptor alpha gene. *Mol Cell Biol.* Oct 1995;15(10):5830-5845.
187. Richer E, Campion CG, Dabbas B, White JH, Cellier MF. Transcription factors Sp1 and C/EBP regulate NRAMP1 gene expression. *FEBS J.* Oct 2008;275(20):5074-5089.
188. Zaragoza K, Begay V, Schuetz A, Heinemann U, Leutz A. Repression of transcriptional activity of C/EBPalpha by E2F-dimerization partner complexes. *Mol Cell Biol.* May 2010;30(9):2293-2304.
189. Porse BT, Pedersen TA, Xu X, et al. E2F repression by C/EBPalpha is required for adipogenesis and granulopoiesis in vivo. *Cell.* Oct 19 2001;107(2):247-258.
190. Motomura Y, Kitamura H, Hijikata A, et al. The transcription factor E4BP4 regulates the production of IL-10 and IL-13 in CD4+ T-cells. *Nat Immunol.* May 2011;12(5):450-459.
191. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science.* Feb 10 2006;311(5762):796-800.
192. Park MJ, Kim HY, Kim K, Cheong J. Homeodomain transcription factor CDX1 is required for the transcriptional induction of PPARgamma in intestinal cell differentiation. *FEBS Lett.* Jan 5 2009;583(1):29-35.
193. Wang D, Paz-Priel I, Friedman AD. NF-kappa B p50 regulates C/EBP alpha expression and inflammatory cytokine-induced neutrophil production. *J Immunol.* May 1 2009;182(9):5757-5762.

194. Reinhold W, Emens L, Itkes A, Blake M, Ichinose I, Zajac-Kaye M. The myc intron-binding polypeptide associates with RFX1 in vivo and binds to the major histocompatibility complex class II promoter region, to the hepatitis B virus enhancer, and to regulatory regions of several distinct viral genes. *Mol Cell Biol.* Jun 1995;15(6):3041-3048.
195. Wouters BJ, Koss C, Delwel R. Gene expression profiling for improved dissection of acute leukemia: a recently identified immature myeloid/T-lymphoid subgroup as an example. *Blood Cells Mol Dis.* May-Jun 2008;40(3):395-400.
196. Wang W, Wang X, Ward AC, Touw IP, Friedman AD. C/EBPalpha and G-CSF receptor signals cooperate to induce the myeloperoxidase and neutrophil elastase genes. *Leukemia.* May 2001;15(5):779-786.
197. Tinel M, Berson A, Elkahwaji J, Cresteil T, Beaune P, Pessayre D. Downregulation of cytochromes P450 in growth-stimulated rat hepatocytes: role of c-Myc induction and impaired C/EBP binding to DNA. *J Hepatol.* Aug 2003;39(2):171-178.
198. Kummalue T, Friedman AD. Cross-talk between regulators of myeloid development: C/EBPalpha binds and activates the promoter of the PU.1 gene. *J Leukoc Biol.* Sep 2003;74(3):464-470.
199. 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.
200. 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.
201. 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.
202. 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.
203. 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.
204. 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.
205. 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.
206. 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.
207. 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.
208. 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.
209. 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.
210. 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.
211. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976;63:655-660.
212. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics.* 1979;6(2):65–70.
213. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biom J.* Feb;52(1):70-84.

214. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society.* 1996;58(1):267-288.
215. 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.
216. 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.
217. Yoshida K, Sanada M, Shiraishi Y, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature.* Oct 6 2011;478(7367):64-69.
218. Visconte V, Makishima H, Maciejewski JP, Tiu RV. Emerging roles of the spliceosomal machinery in myelodysplastic syndromes and other hematological disorders. *Leukemia.* Dec 2012;26(12):2447-2454.
219. 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.* Dec 30 2012.
220. Ogawa S. Splicing factor mutations in myelodysplasia. *Int J Hematol.* Oct 2012;96(4):438-442.
221. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med.* Oct 13 2011;365(15):1384-1395.
222. Thol F, Kade S, Schlarmann C, et al. Frequency and prognostic impact of mutations in SRSF2, U2AF1, and ZRSR2 in patients with myelodysplastic syndromes. *Blood.* Apr 12 2012;119(15):3578-3584.
223. Cazzola M, Rossi M, Malcovati L. Biologic and clinical significance of somatic mutations of SF3B1 in myeloid and lymphoid neoplasms. *Blood.* Jan 10 2013;121(2):260-269.
224. Malcovati L, Papaemmanuil E, Bowen DT, et al. Clinical significance of SF3B1 mutations in myelodysplastic syndromes and myelodysplastic/myeloproliferative neoplasms. *Blood.* Dec 8 2011;118(24):6239-6246.
225. Damm F, Kosmider O, Gelsi-Boyer V, et al. Mutations affecting mRNA splicing define distinct clinical phenotypes and correlate with patient outcome in myelodysplastic syndromes. *Blood.* Apr 5 2012;119(14):3211-3218.
226. Bennett JM, Catovsky D, Daniel MT, et al. Proposals for the classification of the myelodysplastic syndromes. *Br J Haematol.* Jun 1982;51(2):189-199.
227. 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.* Aug 1976;33(4):451-458.
228. Sabattini E, Bacci F, Sagramoso C, Pileri SA. WHO classification of tumours of haematopoietic and lymphoid tissues in 2008: an overview. *Pathologica.* Jun 2010;102(3):83-87.
229. Delwel R, Salem M, Pellens C, et al. Growth regulation of human acute myeloid leukemia: effects of five recombinant hematopoietic factors in a serum-free culture system. *Blood.* Dec 1988;72(6):1944-1949.
230. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* Apr 2003;4(2):249-264.
231. Custom CDFs http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp.
232. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* Aug 15 2009;25(16):2078-2079.
233. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med.* Jul 1990;9(7):811-818.
234. 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 2010;115(3):453-474.
235. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.* 1987;20.
236. Gell D, Kong Y, Eaton SA, Weiss MJ, Mackay JP. Biophysical characterization of the alpha-globin binding protein alpha-hemoglobin stabilizing protein. *J Biol Chem.* Oct 25 2002;277(43):40602-40609.
237. Kihm AJ, Kong Y, Hong W, et al. An abundant erythroid protein that stabilizes free alpha-haemoglobin. *Nature.* Jun 13 2002;417(6890):758-763.
238. Husmeier D. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics.* Nov 22 2003;19(17):2271-2282.
239. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics.* Jun 15 2006;22(12):1540-1542.

240. Ji H, Wong WH. TileMap: create chromosomal map of tiling-array hybridizations. *Bioinformatics.* Sep 15 2005;21(18):3629-3636.
241. Gupta M. Generalized hierarchical markov models for the discovery of length-constrained sequence features from genome tiling-arrays. *Biometrics.* Sep 2007;63(3):797-805.
242. Eichner J, Zeller G, Laubinger S, Ratsch G. Support vector machines-based identification of alternative splicing in Arabidopsis thaliana from whole-genome tiling-arrays. *BMC Bioinformatics.* 2011;12:55.
243. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol.* May 25 1975;94(3):441-448.
244. Young JA, Johnson JR, Benner C, et al. In silico discovery of transcription regulatory elements in Plasmodium falciparum. *BMC Genomics.* 2008;9:70.
245. Ritchie MD, Bush WS. Genome simulation approaches for synthesizing in silico datasets for human genomics. *Adv Genet.* 2010;72:1-24.
246. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. *Br J Pharmacol.* Sep 2007;152(1):21-37.
247. Carvajal-Rodriguez A. Simulation of genes and genomes forward in time. *Curr Genomics.* Mar 2010;11(1):58-61.
248. Carvajal-Rodriguez A. Simulation of genomes: a review. *Curr Genomics.* May 2008;9(3):155-159.
249. El-Maarri O. DNA-methylation and human diseases. *Adv Exp Med Biol.* 2003;544:135-144.
250. Oda M, Greally JM. The HELP assay. *Methods Mol Biol.* 2009;507:77-87.
251. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet.* Mar 2003;33 Suppl:245-254.
252. DeZern AE, Sung A, Kim S, et al. Role of allogeneic transplantation for FLT3/ITD acute myeloid leukemia: outcomes from 133 consecutive newly diagnosed patients from a single institution. *Biol Blood Marrow Transplant.* Sep 2011;17(9):1404-1409.
253. Kuwatsuka Y, Miyamura K, Suzuki R, et al. Hematopoietic stem cell transplantation for core-binding factor acute myeloid leukemia: t(8;21) and inv(16) represent different clinical outcomes. *Blood.* Feb 26 2009;113(9):2096-2103.
254. Marcucci G, Haferlach T, Dohner H. Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *J Clin Oncol.* Feb 10 2011;29(5):475-486.
255. 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.* Jul 12 2012;120(2):395-403.
256. Fasan A, Eder C, Haferlach C, et al. GATA2 mutations are frequent in intermediate-risk karyotype AML with biallelic CEBPA mutations and are associated with favorable prognosis. *Leukemia.* Feb 2013;27(2):482-485.
257. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008;9:387-402.
258. Bellman RE. Dynamic programming *Princeton University Press, ISBN 978-0-691-07951-6.* 1957.

SUMMARY

This thesis consists out of four sections: 1. the creation of a statistical methodology for the detection of candidate regions in the genome using tiling-array or next-generation sequencing technology. 2. Applications of the newly developed method to study DNA-interaction of wild-type and mutant C/EBPα, and to unravel the role of C/EBPα in transformation of cells with myeloid/T-lymphoid features. 3. Analysis of the molecular and clinical behaviour of one specific subtype of human leukemia: patients with mutations in *CEBPA,* the gene encoding C/EBPα. 4. Discovery of specific AML subtypes using combined gene expression and DNA-methylation profiles by bioinformatical approaches in a cohort of human AML.

The first section (Chapter 2 to 4) describes a novel tool to analysis data obtained from tiling-array hybridizations and next-generation sequencing of DNA obtained from, e.g. chromatin immunoprecipitation (ChIP) experiments. These technologies are frequently used to address fundamental biological questions, such as: "*Where does a transcription factor bind to the DNA?", "Can we discriminate active from inactive chromatin in the genome?*" Or "*Can we identify methylated regions in the genome?*". These questions may lead to better understanding of protein-DNA-interactions, gene activation or the identification of important tumour suppressor genes that are repressed as the result of DNA-methylation. By array hybridization or next-generation sequencing millions of signal data-points on the genome are determined. To detect biological relevant regions, signals should be discriminated from non-specific signals (background/ technical variation) to minimize the number of false positives. Candidate regions in the genome are defined by an increase of the signal (read depth or probe intensity) but depends on the type of experiment. There are no standard criteria that define the signals of the candidate regions. So far multiple methods have been developed for the detection of candidate regions but many of these methods are designed for one particular type of experiment and/or require various user defined parameters to model a candidate region, such as maximum bandwidth, the minimum number of data-points in a region or a minimum signal intensity. In Chapter 2 we describe our statistical framework called HAT (Hypergeometric Analysis of Tiling-arrays) to more accurately detect candidate regions that are the result of hybridization of chromatin *immunoprecipitated* DNA-fragments to tiling-arrays. We show that HAT has superior advantages over existing methodologies as it detects candidate regions by specifying only one parameter, the significance level α. We detected candidate regions in various tiling-array experiments such as ChIP-on-chip, MeDIP-on-chip or modified histone pool down experiments.

Although the successes of HAT, we are passing a phase where next-generation sequencing technology is replacing tiling-array technology. In Chapter 3 we describe HATSEQ (Hypergeometric Analysis of Tiling-arrays and Sequence data), which uses the statistical framework of HAT but with substantial improvements so that it can be applied to next-generation chip sequencing data. A comparison with existing ChIP-Seq methodologies showed strong overlap in the detection of candidate regions but our method showed consistently better delineation of the region boundaries and therefore more specificity for the actual binding site. In addition, HATSEQ includes analysis to address the

biological meaning of the detected candidate regions and, includes a graphical-user-interface that lowers the barrier for researcher to analyze their data without the need of scripting languages.

In Chapter 4 we used HAT for the detection of viral integration sites that potentially harbour new tumour suppressor genes in a so called MeDIP-on-chip dataset. This resulted into the identification of a tumour suppressor gene that is associated with outcome in AML. In the second section (Chapter 5 and 6) we studied the molecular mechanisms of action and DNA-binding of C-terminal mutant C/EBPα (Chapter 5) and wild-type C/EBPα (Chapter 6) in a myeloid cell line model using ChIP-on-chip, and by utilizing our previously developed model (HAT). For the latter experiment we detected binding-partners of C/EBPα that are associated with T-cell development. This is especially of interest because it has previously been shown that silencing of *CEBPA* expression levels in AML was negatively correlated with the expression levels of T-cell related genes. We therefore pursued the relation between *CEBPA* expression levels and the downregulated expression levels of T-cells genes. We propose a functional role of C/EBPα among the transforming events that drives the development of mixed myeloid/T-lymphoid leukemia.

In the third section of this thesis (Chapter 7 and 8) we investigated a group of patients carrying a mutation in the gene *CEBPA.* In Acute Myeloid Leukemia (AML), *CEBPA* mutations can roughly be separated into two subgroups, i.e., those with a single mutation (*CEBPA*^{sm}) and those with double mutations (*CEBPA*^{dm}). In Chapter 7 we investigated the clinical outcome and gene expression profiles of *CEBPA*^{dm} versus *CEBPA*^{sm} and the effect of concurrent gene mutations (*NPM1*/ *FLT3*^{ITD}). We show for *CEBPA*^{dm} patients a lower frequency of concurrent mutations compared to *CEBPA*^{sm} patients. The outcome for *CEBPA*^{sm} patients is dominated by concurrent mutations in the genes *NPM1* and *FLT3* (*FLT3*^{ITD}). We report that *CEBPA*^{dm} is an independent prognostic factor for favourable outcome. In support of the prognostic differences between *CEBPA*^{dm} and *CEBPA*^{sm}, striking differences between the gene expression profiles are detected. Patient with *CEBPA*^{dm} expressed a unique gene expression signature that allows further classification/refinement of AML whereas this was not possible for *CEBPA*^{sm} patients. For the group of *CEBPA*^{dm} patients we describe in Chapter 8 the effect of post remission allogeneic and autologous hematopoetic stem cell transplantation (alloHSCT, autoHSCT respectively) compared to *CEBPA*^{dm} patients consolidated with chemotherapy. We show that adults with AML with the distinctive *CEBPA*^{dm} genotype benefit from alloHSCT and autoHSCT in first complete remission (CR) with respect to relapse-free survival (RFS) compared to patients consolidated with chemotherapy. This benefit is not seen in overall survival (OS), apparently due to a high second CR rate after salvage therapy. However, there was a trend for a better OS for alloHSCT, but not for autoHSCT performed in first CR.

In the final section (Chapter 9) we inferred that novel groups of AML can be detected by combining gene expression (GEP) and DNA-methylation profiles (DMP). It has been shown in previous studies that mutations in splice factor (SF) genes occur frequently in Myelodysplastic Syndromes (MDS), including Refractory Anemia with Ring Sideroblasts (RARS), Refractory Anemia with Excess of Blasts (RAEB) or transformation (RAEB(t)). These SF-gene mutations, although less frequently, have also been reported in human acute myeloid leukaemia's (AML). By using both GEP and DMP-dataset we assessed the differences and similarities between MDS and AML with SF-gene mutations. The

combined dataset resulted into an optimal hierarchical clustering containing 18 patient clusters. Among these clusters we identified, besides previously identified AML subgroups also two novel AML subgroups. These two subgroups were not found by using GEP or DMP alone. Both clusters (cluster number 3 and 11) contain patients that are enriched for RAEB(t) cases and splice factor mutations. However there are differences, cluster 3 is enriched for mutations in splice factor *SRSF2* whereas cluster 11 is significantly enriched for mutations in splice factor *U2AF35* and *SF3B1.* Another major difference is the enrichment for *N-RAS* and *K-RAS* mutations in cluster 3. The most discriminative novel cluster 11 presented AMLs with unique erythroid features based on the following findings: 1. Enrichment of pathways associated with erythroid development, when differentially expressed and methylated genes were analysed. 2. High percentages of erythroblasts. 3. Presence of patient samples with a RAEB or a RAEBt. 4. Frequent appearance of ring sideroblasts. Ours is the first report of the clustering of splicing mutants in human neoplasia using both gene expression and DNA-methylation profiles.

SAMENVATTING

Acute Myeloïde Leukemie (AML) is een vorm van leukemie waarbij de uitrijping van bloedcellen op een kwaadaardige manier verandert. Alle bloedcellen worden gemaakt in het beenmerg. In een normale (gezonde) situatie zullen onrijpe hematopoietische stamcellen in het beenmerg via verschillende differentiatie-stadia uitrijpen tot functionele bloedcellen: de rode bloedcellen, witte bloed cellen en bloedplaatjes. Dit proces van bloedcelproductie gebeurt dagelijks en met miljoenen tegelijk. In een abnormale (ongezonde) situatie kunnen deze cellen ongedifferentieerd blijven. De cellen hebben dan geen functie en kunnen vervolgens de normale (gezonde) cellen in het beenmerg en bloed verdringen, met als resultaat een te kort aan normale rijpe bloedcellen. Om patiënten goed te kunnen behandelen moeten we meer weten over de genetische afwijkingen die ten grondslag liggen aan de maligniteit. Er is al veel onderzoek gedaan naar de maligniteit in patiënten met AML, zo is bekend dat AML een heterogene ziekte is, dat wil zeggen dat het niet één ziekte is maar een verzameling van aandoeningen. Het is ook bekend dat AML vaak te wijten is aan genetische afwijkingen. De meest bekende zijn chromosomale veranderingen zoals inv(16), of herschikkingen van de chromosomen zoals gebeurt bij translocatie t(15;17) en t(8;21). Hierdoor kunnen delen van twee losse genen bij elkaar gebracht worden waardoor vervolgens een nieuw fusie-gen ontstaat. Andere bekende afwijkingen in AML zijn subtiele mutaties in kanker-kritische genen, zoals “Internal Tandem Duplications” in *FLT3* (*FLT3ITD*) of mutaties in nucleophosmin (*NPM1*) of "CCAAT enhancer binding protein alpha" (*CEBPA*). Verder is bekend dat genetische afwijkingen gerelateerd zijn met prognose. Ondanks de vele verrichtte studies en de daardoor verkregen inzichten, worden nog steeds nieuwe genetische afwijkingen in patiënten met AML gevonden. Sub classificatie van patiënten zal in de toekomst dan ook toenemen.

In dit proefschrift wordt de rol van C/EBPα in myeloïde leukemie nader onderzocht en in het bijzonder bij één specifieke groep van patiënten; namelijk die met een afwijking in het gen dat codeert voor C/EBPα. C/EBPα is één van de master regulatoren voor myeloïde differentiatie. Een gemuteerd *CEBPA* codeert voor een niet of deels functioneel C/EBPα-eiwit dat effect heeft op de cel differentiatie. Daarnaast wordt er in dit proefschrift een statistische methode beschreven dat ontwikkeld is om genoom-breed eiwit-DNA interacties te detecteren met behulp van zogenaamde tiling-array en next-generation sequencing technologie (NGS). De methode is vervolgens gebruikt om virale integratie-sites te detecteren die mogelijk samenhangen met nieuwe tumor suppressor genen. Verder werd de technologie toegepast om (gemuteerde) C/EBPα-DNA interacties te identificeren en te associëren met gen-expressie. Daarbij werd de rol van C/EBPα in myeloïde/T-lymfoïde transformatie geanalyseerd. Verder werd voor deze groep van patiënten het effect van postremissie allogene en autologe hematopoietische stamceltransplantatie onderzocht. Tot slot zijn er nieuwe AML subgroepen geïdentificeerd door gen-expressie data en DNA methylatie data te combineren.

In de eerste sectie (hoofdstuk 2, 3 en 4) wordt een statistische methode beschreven die we hebben ontwikkeld om data te analyseren die verkregen zijn met behulp van tiling-array of next-generation sequencing technologie. Deze technieken kunnen worden gebruikt om fundamentele biologische vragen te beantwoorden, zoals: “*Waar binden*

transcriptie factoren op het DNA?", "Kunnen we actief en inactief chromatine op het genoom van elkaar onderscheiden?" of "*Kunnen we gemethyleerde gebieden identificeren in het genoom?*". Het beantwoorden van deze vragen kan leiden tot een beter begrip van bijvoorbeeld eiwit-DNA interacties, gen activering of de identificatie van belangrijke tumor suppressor genen die onderdrukt worden als gevolg van DNA-methylering. Door gebruik te maken van array hybridisatie of next-generation sequencing kunnen miljoenen datapunten genoom-breed gemeten worden. De signalen die kandidaat gebieden representeren moeten vervolgens onderscheiden worden van niet specifieke signalen (achtergrond/technische variatie). De kandidaat gebieden worden gedefinieerd door een toename van het signaal, maar er zijn geen definities die a priori beschrijven wat een kandidaat gebied is (zoals bijvoorbeeld de grootte of sterkte van het signaal). Dit kan namelijk verschillen tussen de verschillende typen experimenten. Tot dusver zijn er verschillende methoden ontwikkeld voor de detectie van kandidaat gebieden. Sommige van deze methodieken zijn ontworpen voor één bepaald type experiment en kunnen meerdere instellingen vereisen om een kandidaat gebied te beschrijven, zoals de minimale sterkte van het signaal, maximale breedte van het kandidaat gebied en minimaal aantal datapunten dat voor een kandidaat gebied moet worden gemeten. In hoofdstuk 2 wordt een door ons ontwikkelde methode, genaamd HAT (Hypergeometrische Analyse van Tiling-arrays) beschreven om kandidaat gebieden te identificeren. We laten zien dat HAT voordelen heeft ten opzichte van andere bestaande methodieken. HAT kan namelijk kandidaat gebieden detecteren in verschillende type tiling-array experimenten, zoals ChIP-on-chip en MeDIP-on-chip, zonder extra kennis nodig te hebben over de te verwachte signaalsterkte/ grootte van het kandidaat gebied. Alhoewel HAT succesvol is toegepast in tiling-array data, passeren we een fase waarin next-generation sequencing technologie de tiling-array technologie gaat, zo niet reeds heeft vervangen. In hoofdstuk 3 beschrijven we HATSEQ (Hypergeometrische Analyse van Tiling-arrays en Sequence data), gebaseerd op de statistische methode die in HAT gebruikt is. HATSEQ heeft aanzienlijke uitbreidingen zodat het ook kan worden toegepast op de nieuwe generatie ChIP-sequence data. Als we de resultaten van HATSEQ vergelijken met die van andere methoden, dan zien we een grote overlap in de gedetecteerde kandidaat gebieden. Desondanks zijn de gebieden gedetecteerd door HATSEQ specifieker. Bovendien kan HATSEQ de gedetecteerde kandidaat gebieden bestuderen op hun eventuele biologische betekenis, zoals de aanwezigheid van motieven en/of verrijking voor specifieke pathways. HATSEQ is overigens ontwikkeld met een grafische gebruikers interface zodat onderzoekers zonder kennis van programmeertalen de methode kunnen toepassen. In hoofdstuk 4 laten we in een MeDIP-on-chip experiment zien dat virale integratie-sites, die mogelijk samenhangen met nieuwe tumor suppressor genen, gedetecteerd kunnen worden. We hebben een tumor suppressor-gen geïdentificeerd dat geassocieerd is met de overleving van patiënten met AML.

De tweede sectie van mijn proefschrift (hoofdstuk 5 en 6) bevat onderzoek waarbij we de functionele rol van de transcriptie factor C/EBPα in myeloïde cel line modelsystemen hebben bestudeerd, door gebruik te maken van ChIP-on-chip technologie. We hebben daarvoor gebruik gemaakt van de door ons eerder ontwikkelde methode, HAT. In hoofdstuk 5 bestuderen we de "binding-sites" van C-terminal mutant C/EBPα en in hoofdstuk 6 is de moleculaire functie van C/EBPα onderzocht. In de loop der jaren is veel inzicht verkregen in de rol van C/EBPα in hematopoietische ontwikkeling, maar de directe "binding-sites" en de associatie met veranderingen in mRNA expressie is nog

grotendeels onbekend. Zo laten we in hoofdstuk 6 zien dat we een groep van genen gedetecteerd hebben waar C/EBPα in staat was om te binden aan promotor sequenties, en de expressie van de gerelateerde genen verlaagd was in een sub groep van AML patiënten, de zogenaamde groep van *CEBPA*-silenced AML patiënten. Deze experimenten ondersteunen de hypothese dat C/EBPα niet alleen kan fungeren als een transcriptionele activator maar ook als een repressor van genen. Onze experimenten lijken er vooral op te wijzen dat expressie van genen, die van belang zijn bij ontwikkeling van T-lymfocyten en waarvan de promotor een interactie aangaat met C/EBPα, geremd wordt door deze transcriptie factor. Dit is vooral van belang omdat al eerder aangetoond was dat het volledig uitschakelen van *CEBPA,* geassocieerd was met de expressie van T-cel gerelateerde genen.

In de derde sectie van dit proefschrift (hoofdstuk 7 en 8) is er onderzoek gedaan aan een groep van AML patiënten met mutaties in het *CEBPA* gen. AML patiënten met *CEBPA* mutaties kunnen kortweg in twee subgroepen worden onderverdeeld, namelijk die met een enkele mutatie (*CEBPA*^{sm}) en die met dubbele mutaties (*CEBPA*^{dm}). Bij *CEBPA*^{dm} patiënten zijn meestal beide allelen betrokken, zodat in de leukemie cellen van deze patiënten geen normaal C/EBPα aanwezig is. In hoofdstuk 7 wordt de prognostische waarde en de gen-expressieprofielen van *CEBPA*^{dm} versus *CEBPA*^{sm} onderzocht. Verder is de rol van andere gen mutaties, namelijk *NPM1* mutaties en *FLT3ITD* bestudeerd. We tonen aan dat *NPM1* mutaties en of *FLT3ITD* vaker voor komen bij *CEBPA*^{sm} patiënten dan bij *CEBPA*^{dm} patiënten met AML. Vervolgens laten wij zien dat *CEBPA*^{dm} een onafhankelijke prognostische factor is voor een gunstige uitkomst. Voor *CEBPA*^{sm} patiënten geldt dat de prognostische waarde wordt gedomineerd door mutaties in *NPM1* en/of *FLT3ITD*. Tot slot zijn er in de gen-expressieprofielen opvallende verschillen gedetecteerd. Patiënten met *CEBPA*^{dm} kunnen worden herkend aan een uniek gen-expressieprofiel. Dit bleek niet mogelijk voor de *CEBPA*^{sm} groep. In hoofdstuk 8 bestuderen we het effect van postremissie allogene en autologe hematopoietische stamceltransplantatie (alloHSCT, autoHSCT respectievelijk) in *CEBPA*^{dm} patiënten. De effecten van de HSCT worden vergeleken met *CEBPA*^{dm} patiënten die alleen chemotherapie hebben gehad. We tonen aan dat *CEBPA*^{dm} patiënten profiteren van alloHSCT en autoHSCT in de eerste complete remissie met betrekking tot “relapse-free survival” in vergelijking met patiënten die alleen chemotherapie gehad hebben.

In de vierde en laatste sectie (hoofdstuk 9) laten we zien dat nieuwe patiënt groepen in AML gedetecteerd kunnen worden door gen-expressie (GEP) en DNA methylatie profielen (DMP) te combineren. In eerdere studies was aangetoond dat mutaties in splice factor (SF) genen voorkomen bij myelodysplastische syndromen (MDS), met refractory anemia en ring sideroblasts (RARS), en met refractory anemia met excess of blasts (RAEB) of in transformatie (RAEBt). Deze splice factor gen mutaties zijn ook gedetecteerd in AML maar in mindere mate. Door gebruik te maken van zowel GEP en DMP-datasets konden we de verschillen en overeenkomsten tussen RAEB, RAEBt en AML met splice factor gen mutaties analyseren. Met de gecombineerde GEP/DMP-dataset hebben we een optimale hiërarchische clustering met 18 clusters verkregen. De clusters bevatten, naast de eerder gevonden AML subgroepen ook twee nieuwe AML subgroepen. Deze twee subgroepen werden niet gevonden door gebruik te maken van alleen de GEP of alleen de DMP-dataset. De twee nieuwe clusters (cluster nummer 3 en 11) bevatten patiënten

die verrijkt zijn voor RAEB, RAEBt en AML met SF-mutaties. Er zijn echter verschillen tussen clusters: cluster 3 is verrijkt voor mutaties in splice factor *SRSF2* terwijl cluster 11 verrijkt is voor mutaties in splice factor *U2AF35* en *SF3B1*. Een ander belangrijk verschil is dat patiënten in cluster 11 unieke erythroid eigenschappen hebben. Dit is bepaald op basis van de volgende bevindingen: 1. verrijking voor gen-pathways die geassocieerd zijn met erythroid ontwikkeling, 2. hoge percentages erythroblasten, 3. het veel voorkomen van RAEB of RAEBt patiënten, en 4. het veel voorkomen van ring sideroblasts. In tegenstelling zien we deze erythroid verrijking niet in de patiënten van cluster 3. In dit cluster zijn *N-RAS* en *K-RAS* mutaties vaak aanwezig. Deze mutaties zijn niet gevonden in cluster 11. Met deze studie laten wij als eerste zien dat nieuwe subgroepen in AML geïdentificeerd en beter gekarakteriseerd kunnen worden door gen-expressie en de DNA-methylatie profielen te combineren.

ABBREVIATIONS

TSS	Transcription Start Site
AlloHSCT	Allogeneic haematopoietic stem cell transplantation
AML	Acute myeloid leukemia
AutoHSCT	Autologous haematopoietic stem cell transplantation
BM	Bone marrow
bZIP	Basic leucine zipper
C/EBPα	CCAAT/enhancer binding protein alpha (protein)
CBF	Core-binding factor
CEBPA	CCAAT/enhancer binding protein alpha (gene)
CEBPA^{dm}	*CEBPA* double-mutation
CEBPA^{sm}	*CEBPA* single-mutation
ChIP	Chromatin immunoprecipitation
ChIP-on-chip	Chromatin immunoprecipitation followed by analysis on array
ChIP-Seq	Chromatin immunoprecipitation followed by sequencing
CID	Cumulative incidence of death
CIR	Cumulative incidence of relapse
CMP	Common myeloid progenitor
CN	Cytogenetical Normal
CR(1)	Complete Remission (First)
dHPLC	Denaturing high performance liquid chromatography
DMP	DNA-methylation profiling
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
E2	β-estradiol
EFS	Event-free survival
ER	Estrogen receptor
ERG	v-ets erythroblastosis virus E26 oncogene homolog (gene)
ETO	Eight twenty one (gene)
EVI1	Ecotropic viral integration site 1 (gene)
FAB	French American British
FDR	false discovery rate
FLT3	FMS-like tyrosine kinase 3 (gene)
FWER	Family-wise error rate
G-CSF	Granulocyte colony-stimulating factor
GEO	Gene expression omnibus
GEP	Gene expression profiling
Gr1.4 MuLV	Graffi 1.4 murine leukemia virus
GSEA	Gene set enrichment analysis
HAT	Hypergeometric analysis of tiling-arrays
HATSEQ	Hypergeometric analysis of tiling-arrays and sequence data

HELP	HpaII tiny fragment enrichment by ligation mediated PCR
HR	Hazard Ratio
HSC	Hematopoietic stem cell
IL-3 / -6	Interleukin 3 / -6
Indel	Insertion or deletion
IP	Immunoprecipitation
iPCR	Inverse PCR
ITD	Internal tandem duplication
MAS	MicroArray Suite
MDS	Myelodysplastic syndrome
MeDIP	Myelodysplastic DNA immunoprecipitation
MeDIP-on-chip	Myelodysplastic DNA immunoprecipitation followed by analysis on array
methyl-seq	DNA-methylation profiling by deep-sequencing
mRNA	Messenger RNA
mRNA-seq	mRNA profiling by deep-sequencing
mVIS	Methylated viral integration site
MYH11	Myosin, heavy chain 11 (gene)
NOTCH1	Notch homolog 1 (gene)
NPM1	Nucleophosmin (gene)
OS	Overall survival
PCR	Polymerase chain reaction
PTP4A3	Protein tyrosine phosphatase type IVA, member 3
PWM	Position Weight Matrices
RAEB(t)	Refractory anemia with excess blasts (in transformation)
RFS	Relapse-free survival
RIM	Retroviral integration mutagenesis
RMA	Robust multi averaging
RNA	Ribonucleic acid
ROI	Region-of-interest
RQ-PCR	Quantitative real-time reverse transcription PCR
RT-PCR	Reverse transcription PCR
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
STAT	Signal transducer and activator of transcription
TAD	Transactivation domain
TCR	T-cell receptor
TFBS	Transcription Factor Binding Sites
WBC	White blood cell
WES	Whole exome sequencing
WGS	Whole genome sequencing
WHO	World Health Organization

PUBLICATIONS

1. **Taskesen E**, Wouters BJ, Avellino R, AlberichJorda M, Tenen DG, Ridder J, Valk PJM, Erpelinck E, Reinders MJT, Delwel R, *A repressor function of C/EBPα uncovered by combining gene expression profiling in AML and chromatin immunoprecipitation in a myeloid differentiation model. In preparation*

2. **Taskesen E**, Havermans M, Lom K, Sanders M, Norden Y, Bindels E, Hoogenboezem R, Reinders MJT, Figueroa ME, Valk PJM, Löwenberg B, Melnick A, 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. *Under review*

3. **Taskesen E**, Wouters B, Delwel R, *HAT: A Novel Statistical Approach to Discover Functional Regions in the Genome*. Springer Series (book Chapter). doi: 10.1007/978-1-62703-607-8

4. **Taskesen E**, Hoogeboezem R, Delwel R, Reinders MJT. (2013) *HATSEQ: Detection and interpretation of peaks in tiling-array and sequence data.* Advances and Applications in Bioinformatics and Chemistry. *In Press*

5. **Taskesen E***, Schlenk RF*, van Norden Y, Krauter J, Ganser A, Bullinger L, Gaidzik VI, Paschka P, Corbacioglu A, Gohring G, Kundgen A, Held G, Gotze K, Vellenga E, Kuball J, Schanz U, Passweg J, Pabst T, Maertens J, Ossenkoppele GJ, Delwel R, Dohner H, Cornelissen JJ, Dohner K, Lowenberg B (2013) The value of allogeneic and autologous hematopoietic stem cell transplantation in prognostically favorable acute myeloid leukemia with double-mutant CEBPA. BLOOD. doi:10.1182/blood-2013-05-503847

6. Bindels EM, Havermans M, Lugthart S, Erpelinck C, Wocjtowicz E, Krivtsov AV, Rombouts E,Armstrong SA, **Taskesen E**, Haanstra JR, Beverloo HB, Dohner H, Hudson WA, Kersey JH, Delwel R, Kumar AR (2012) *EVI1 is critical for the pathogenesis of a subset of MLL-AF9-rearranged AMLs.* BLOOD 119 (24):5838-5849. doi:blood-2011-11-393827

7. Zebisch A, Wolfler A, Fried I, Wolf O, Lind K, Bodner C, Haller M, Drasche A, Pirkebner D, Matallanas D, Rath O, Blyth K, Delwel R, **Taskesen E**, Quehenberger F, Kolch W, Troppmair J, Sill H (2012) *Frequent los of RAF kinase inhibitor protein expression in acute myeloid leukemia. Leukemia* 26 (8):1842-1849. doi:leu201261

8. Horos R, Ijspeert H, Pospisilova D, Sendtner R, Andrieu-Soler C, **Taskesen E**, Nieradka A, Cmejla R, Sendtner M, Touw IP, von Lindern M (2012) *Ribosomal deficiencies in Diamond- Blackfan anemia impair translation of transcripts essential for differentiation of murine and human erythroblasts.* BLOOD 119 (1):262-272. doi:blood-2011-06

9. Beekman R, Valkhof M, Erkeland SJ, **Taskesen E**, Rockova V, Peeters JK, Valk PJ, Lowenberg B, Touw IP (2011) *Retroviral integration mutagenesis in mice and comparative analysis in human AML identify reduced PTP4A3 expression as a prognostic indicator.* PLoS One 6 (10):e26537. doi:10.1371/journal.pone.0026537

10. Meenhuis A, van Veelen PA, de Looper H, van Boxtel N, van den Berge IJ, Sun SM, **Taskesen E**, Stern P, de Ru AH, van Adrichem AJ, Demmers J, Jongen-Lavrencic M, Lowenberg B, Touw IP, Sharp PA, Erkeland SJ (2011) *MiR-17/20/93/106 promote hematopoietic cell expansion by targeting sequestosome 1-regulated pathways in mice.* BLOOD 118 (4):916-925. doi:blood- 2011-02-336487

11. Smith LL, Yeung J, Zeisig BB, Popov N, Huijbers I, Barnes J, Wilson AJ, **Taskesen E**, Delwel R, Gil J, Van Lohuizen M, So CW (2011) *Functional crosstalk between Bmi1 and MLL/Hoxa9 axis in establishment of normal hematopoietic and leukemic stem cells.* Cell Stem Cell 8 (6):649-662. doi:S1934-5909(11)00224-4

12. **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 PJ, Dohner K (2011) *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 (8):2469-2475. doi:blood-2010-09-307280

13. **Taskesen E**, Beekman R, de Ridder J, Wouters BJ, Peeters JK, Touw IP, Reinders MJ, Delwel R (2010) *HAT: hypergeometric analysis of tiling-arrays with application to promoter-GeneChip data.* BMC Bioinformatics 11:275. doi:1471-2105-11-275

*These authors corresponded equally