Gene expression profiling of acute myeloid leukemia

Roel G.W. Verhaak

Gene expression profiling of acute myeloid leukemia

Genoom-brede gen expressie studies aan acute myeloïde leukemie

Proefschrift

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

Prof.dr. S.W.J. Lamberts

en volgens het besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

vrijdag 24 november 2006 om 11:00 uur

door

Roel George Willehad Verhaak

geboren te Wijchen

Promotiecommissie

Promotor	Prof. dr. B. Löwenberg
Overige leden	Prof. dr. R. Pieters Dr. L.J. Van 't Veer Prof. dr. P.J. Van der Spek
Co-promotor	Dr. P.J.M. Valk

The work described in this thesis was performed at the Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands. This work was supported by grants from the Dutch Cancer Society (Koningin Wilhelmina Fonds) and the Erasmus Medical Center (Revolving Fund).

The printing of this thesis was financially supported by Skyline Diagnostics BV and Affymetrix Ltd.

Cover design by Roel G.W. Verhaak
ISBN 90-8559-238-0

Contents

Omnes una manet nox

Chapter 1

Introduction

1. Introduction

Acute myeloid leukemia (AML) is a highly heterogeneous disease which it is thought to be the result of a multi-step process of genetic transformations. This has a large impact on the clinical presentation of the disease, and on the development, the choice and management of therapy of patients with AML. This thesis describes various aspects of AML in the context of genome-wide expression profiling, a promising new technique for new diagnostic approaches and therapeutic target discovery in AML.

1.1 Hematopoiesis

The formation of mature blood cells, or hematopoiesis, is characterized by the existence of pluripotent hematopoetic stem cells (HSC), which upon stimulation with different growth factors proliferate and differentiate into all functional end cells of the hematopoietic system (Figure 1).

1.1.1 Embryonal development of the hematopoietic system

In the first phase post-conception, the fertilized egg proliferates and forms a blastocyst, a thin-walled hollow sphere made up of an outer layer of cells, a fluid filled cavity and an inner cell mass containing pluripotent stem cells. In two weeks, blastocyst cells differentiate into three different functional germ layers, ectoderm, endoderm and mesoderm. Like muscle and central nervous system tissue, the hematopoietic tissues arise from the mesoderm layer. The hematopoietic system is established in a sequential way during development (Figure 2). In mice, hematopoietic stem cells or HSC can be found in the yolk sac as early as day E7. Until recently, these HSC were thought to originate from sites in the yolk sac (1). The hematopoietic system of birds and amphibians, however, was shown to arise from an area surrounding the dorsal aorta, gonads and pro-/mesonephros (AGM) (2, 3) and this observation was later on reproduced in vertebrates. It is thought that yolk sac hematopoietic cells, i.e. primitive hematopoiesis, are transient blood cells only present during embryogenesis, whereas definitive hematopoiesis arises from the AGM-region (4-7). There are indications that HSCs colonizing the fetal liver are partially generated in the placenta (8). Later in embryonic development, HSC from the AGM colonize the liver and subsequently the bone marrow, which will remain the site of hematopoiesis during adult life (9).

1.1.2 Hematopoietic stem cells and hematopoiesis

All mature and functional end cells of the hematopoietic system, i.e. cells of bone marrow, blood, spleen and thymus, are derived, following proliferation, differentiation and maturation, from hematopoietic stem cells. This process is under tight control of a network of proliferation and differentiation stimulating glycoproteins such as granulocyte macrophage-colony stimulating factor , KIT ligand and erythropoetin (Figure 1) (10-16). In addition, HSC are tightly coupled with stromal and mesenchymal cells that constitute and produce the microenvironment. It is this microenvironment that produces a variety of stimulatory and inhibitory factors that control and determine HSC development (17, 18).
Various progenitor and differentiated blood cell types are derived from the HSC and can be classified into two major branches; a myeloid branch, giving rise to

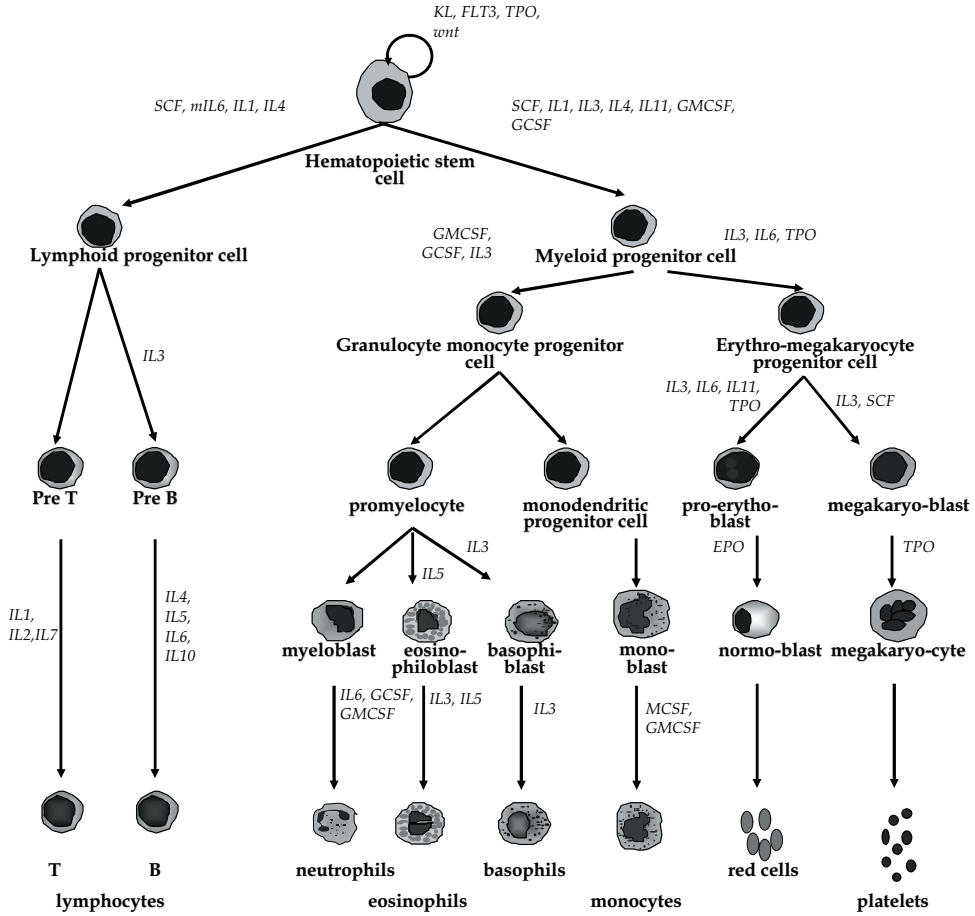


Figure 1. Schematic representation of hematopoiesis. *Different types of mature blood cells arise from the hematopoietic stem cells, passing through various different stages of progeny. These processes require distinct growth factors including different interleukines (IL), thrombopoeitin (TPO), granulocyte colony stimulating factor (GCSF) and macrofage colony stimulating factor (MCSF).*

granulocytes (neuthrophiles, eosinophils and basophils), monocytes/macrophages, erythrocytes and platelets; and a lymphoid branch from which B-cells and T-cells are derived (Figure 1).

Pluripotency is a key feature of HSC. An additional criterium to qualify as true HSC is self-renewal capacity (10, 19, 20). HSC have a very high proliferation capacity; in fact, one cell can sustainably repopulate the entire hematopoietic system in mice in three months (21-23). Several marker proteins, such as c-KIT, CD34, macrophage antigen 1 and stem cell antigen 1, can be used to establish the origin of different

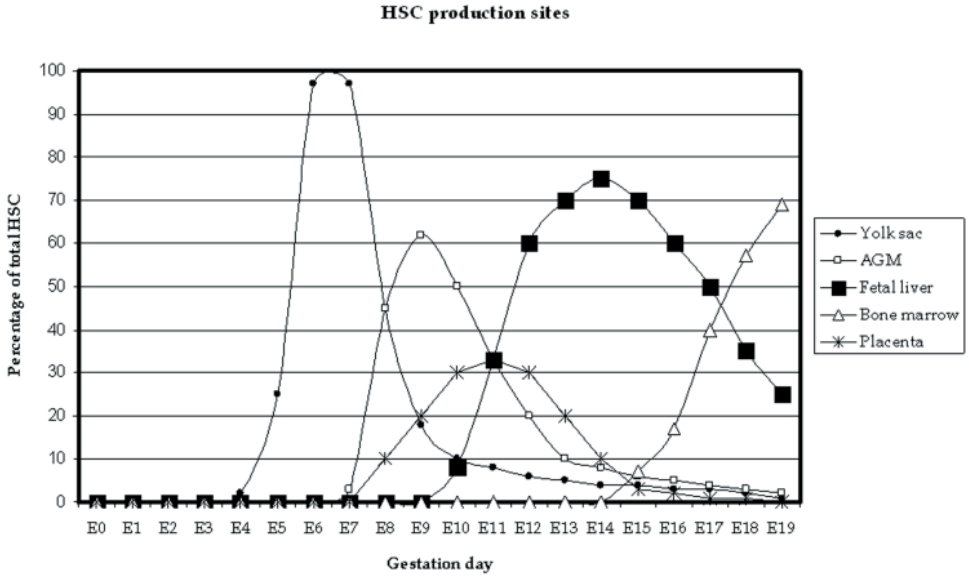


Figure 2. Schematic representation of the origin of HSC in different stages of development. *During development, hematopoiesis shifts from sites in the yolk sac to ultimately the bone marrow.*

HSCs (24). Stem cell antigen 1 has been identified as a marker for AGM HSC, where additional macrophage antigen 1 presence may predispose for liver colonization (25, 26). Using markers, such as CD34, stem cell antigen 1, c-KIT and absence of LIN1, 0.1% of hematopoietic cells have been estimated as HSC (15).

1.2 Acute myeloid leukemia

Leukemia (derived from the Latin terms leuko (white) and heme (blood)) is a disease in which the hematopoietic system completely or partially fails to produce functional blood cells. In general, leukemia can be distinguished into myeloid leukemia and lymphoid leukemia, which relates to the type of progenitor cell from which the malignant cells are derived. Both can be further subdivided into acute and chronic forms. In chronic disease increased numbers of more mature malignant cells, which are partially functional, are present in the bone marrow and blood. In acute leukemia, immature blood cells accumulate in bone marrow and blood, which leads to complete disruption of normal hematopoiesis.
In the United States, incidence of AML is approximately 3 cases per 100.000 inhabitants per year. This increases with age, up to 12.6 cases per 100.000 inhabitants older than 65 per year (27). Incidence in the Netherlands maybe somewhat less, with 726 new cases in 1996, which translates to 1.8 new cases per 100.000 inhabitants per year (28). With 65.000 people diagnosed with cancer annually, this makes AML one of the less frequently found cancers in the Netherlands (29).

1.2.1 Diagnosis and prognosis

AML diagnosis and establishing prognosis is a multidisciplinary process in which morphology, cytology, immunophenotyping, cytogenetics and molecular

diagnostics are combined. The primary diagnosis of AML is established through cytological and cytochemical identification of leukemic blast cells in blood and bone marrow. Further, the myeloid lineage is confirmed by immunophenotyping. These examinations serve as the basis for the French-American-British classification (30-33)., which does not harbor prognostic value. Recently, the French-American-British classification has been replaced by the World Health Organization classification (34, 35). The latter classification also incorporates cytogenetic and molecular analysis and uses clinical features e.g. preceding therapy or an antecedent hematological disease. The World Health Organization classification has recently replaced the French-American-British classification .

Immunophenotyping is used to characterize the blast population and plays a role mainly in diagnosis rather than in prognosis. Using several monoclonal antibodies, such as CD34 (stem/progenitor cell marker), CD2, CD3 and CD5 (T-lymphoid markers), CD19, CD20 (B-lymphoid markers), myeloperoxidase, CD14, CD15 (monoblast/monocyte/macrophage markers) and glycophorin A (erythrocyte marker), cellular differentiation lineage of the leukemia is established.

More traditional prognostic markers are known, e.g. age and white blood cell count per mm3 (36). The classification of prognostic risk is commonly derived from the presence or absence of particular cytogenetic and molecular markers. For example, AMLs with recurrent balanced translocations such as inversion(16), translocation(15;17), translocation(8;21) are associated with a relatively favorable prognosis. In addition to the recurrent translocations, t(8;21), inv(16) and t(15;17), several other acquired deletions and amplifications, such as deletions of chromosome 5 or 7, or trisomy of chromosome 8 are frequently present in patients with AML (37-46).

With remission induction therapy, 70-80% of young and middle-aged adults will successfully obtain a complete remission. On average, overall survival in AML is approximately 40% after 5 years for patients less than 60 years of age. Based on presence of inversion(16), translocation(15;17) or translocation(8;21), approximately 20 percent of de novo AMLs is classified as favorable prognosis. Patients with these abnormalities show an overall survival after 5 years of 60 to 70% and a complete remission rate of 90% (43). Presence of translocation(9;22), translocation(6;9), deletion(5) or (5q), deletion(7) or (7q) or translocation(11q23), represents another 20 percent of patients, who are usually classified as having an unfavorable prognosis. Overall survival in these cases is approximately 15% (47, 48).

Translocations of specific chromosomes are not only identified by cytogenetics, but the resulting fusion mRNA transcripts are also detectable by molecular diagnostics. Molecular diagnostic techniques, such as Southern blotting and reverse transcriptase polymerase chain reaction, offer the advantage of relatively easy implementation, the possibility to detect cryptic chromosomal abnormalities as well as the advantage of measuring abnormal subpopulations with increased sensitivity. In addition to the detection of recurrent translocations, molecular diagnostics is also applied to detect genetic point mutations and small duplications and deletions (49-52).

Not only large chromosomal aberrations have prognostic impact; several genetic mutations have been shown to impact on prognosis have been identified. Mutations in the gene encoding CCAAT/enhancer binding protein alfa (CEBPα) define AML with a relatively good response to treatment (51, 52) , while internal tandem

duplications and tyrosine kinase duplications in fms-related tyrosine kinase 3 gene (*FLT3*) are associated with a poor prognosis (53, 54). Other mutations, such as in the genes encoding tumor protein 53 (*TP53*), and Wilms'tumor suppressor (*WT1*) and high expression of the brain and acute leukemia cytoplasmic gene (*BAALC*), *WT1* and ecotropic virus integration 1 (*EVI1*) are considered markers for unfavorable prognosis (55-57). An overview of prognostic markers and their prevalence is given in Table 1.
Although not routinely applied at most institutions, autonomous growth of blast cells has also been shown to have prognostic impact (58, 59).

1.2.2 Treatment

Survival rates for AML have increased during recent decennia. Currently, the 5-years survival rate for AML patients under the age of 60 is approximately 40 percent, coming from a value of only 15 percent in the 1970s (60).
The first goal of AML treatment is to induce a remission in order to restore normal hematopoiesis. The second goal of treatment following the initial effort of leukemia eradication is to prevent relapse. This second phase of treatment may involve the use of additional chemotherapy or the application of hematopoietic stem cell transplantation.
A special case is the treatment of acute promyelocytic leukemia, a subclass of acute myeloid leukemia. Acute promyelocytic leukemia responds uniquely to the differentiation-inducing effects of trans-retinoic acid, usually administered in combination with chemotherapy. Retinoic acid is in fact a ligand of the retinoic acid receptor alfa protein, which is involved in the acute promyelocytic leukemia specific t(15;17) chimeric fusion protein (61).
New developments in drug therapeutics include the development of small molecules targeting specific enzymes. An example of such a small molecule is imatinib, a drug originally developed for the treatment of chronic myeloid leukemia (62). Over 90% of chronic myeloid leukemia cases presents with a reciprocal translocation of chromosomes 9 and 22. Imatinib specifically inhibits the Abelson murine leukemia tyrosine kinase, which is part of the fusion protein resulting from the t(9;22) translocation. Imatinib binds to the centrally located activation loop of the Abelson murine leukemia protein (63). This translocation is also infrequently (1-2%) seen in AML, thus in special cases of AML, imatinib may be applied. In AML, other small molecules are being developed, e.g. those that target FLT3 and RAS (64, 65).

1.2.3 Pathogenesis

AML is a highly heterogeneous stem cell disease in which a variety of cytogenetic aberrations and molecular mutations can be involved. These abnormalities relate to different pathogenetic mechanisms affecting proliferation, differentiation,apoptosi s, self-renewal and DNA repair. AML is not initiated by a single genetic event but it is thought to be the result of a multi-step process in which genetic transformations accumulate (66-68).
Many routes of leukemogenesis are yet to be elucidated. During many years of scientific research the structural and functional consequences of balanced reciprocal translocations have been investigated. In AML, chromosomal translocations lead to the formation of chimeric fusion proteins. Examples of genes that generate fusion

Cytogenetic abnormality	Genes involved	Prognosis	Frequency
t(8;21)(q22;q22)	*AML1-ETO*	Favourable	8-12%
inv(16)inv(16)(p11;q22)	*CBFβ-MYH11*	Favourable	4-10%
t(15;17)(q22;q11)	*PML-RARa*	Unfavourable	8-12%
11q23 abnormalities	*MLL*	Unfavourable	4-6%
-5/5q	Unknown	Unfavourable	2-4%
-7/7q	Unkown	Unfavourable	6-8%
t(6;9)(q34;q11)	*DEK-CAN*	Unfavourable	1%
t(9-22)(q34;q11)	*BCR-ABL*	Unfavourable	<1%
3q26 abnormalities	*EVI1*	Unfavourable	1-3%
3q26 abnormalities	Unknown	Unfavourable	4-6%
+8	Unknown	Unclear	8-10%

Table 1A. Cytogenetic abnormalities in AML

Marker	Prognosis	Frequency
FLT3 internal tandem duplication	Unfavourable	25%
FLT3 tyrosine kinase domain	Unclear	10-12%
High *EVI1* expression	Unfavourable	5-10%
TP53 mutation	Unfavourable	3-5%
High *BAALC* and *WT1* expression	Unfavourable	30-40%
N-RAS mutation	Unclear	8-12%
K-RAS mutation	Unclear	3-5%
High *ERG* expression	Unfavourable	25%
CEBPa	Favourable	5-10%

Table 1B. Molecular markers added to cytogenetics with prognostic significance

proteins are the *PML-RARa* gene (promyelocytic leukemia - retinoic acid receptor alfa), the *AML1-ETO* gene (acute myeloid leukemia – eight twenty-one) and the *CBFβ–MYH11* gene (core binding factor beta - myosin heavy chain 11) in AMLs with t(15;17), t(8;21) and inv(16), respectively (69-72).

PML, part of the PML-RARα fusion protein, in normal cells is present in nuclear structures called PML nuclear bodies. PML is involved in several processes, such as regulation of cyclin D1 via interaction with eukaryotic translation initiation factor 4E, as a tumor suppressor protein through regulation of nuclear body proteins and is also believed to have a role in TP53-dependent as well as TP53-independent apoptosis pathways (73-78). PML has also been proposed to function as a transcriptional regulator via interaction with pRb in the nuclear bodies, thereby displaying an inhibitory effect on the pRb-regulated activation of the glucocorticoid receptor (79), and its interaction with Pu.1, leading to transcriptional repression of the Pu.1-dependent epidermal growth factor receptor gene promoter. Decreased expression of Pu.1 has been shown to induce AML in mice (80). However, there

has been no direct experimental evidence supporting a direct effect of PML on transcriptional regulation, and it is important to note that PML nuclear bodies do not associate with sites of active transcription and do not localise with nascent DNA (81). Under normal physiological conditions, RARα dimerizes with retinoic-X-receptors to form a complex that act as nuclear retinoid receptors (82). The PML-RARα fusion protein is a dominant negative RARα mutant, indicating that it can bind to retinoic-X-receptors with stronger affinity than the wild-type (83) and it has been shown that in absence of retinoic acid, the complex can repress transcriptional activation through histone deacetylation (84, 85). It is therefore thought that oncogenic retinoic acid receptors mediate leukemogenesis through aberrant chromatin acetylation. However, the PML-RARα fusion in itself is not sufficient to cause leukemia in genetically modified mice, indicating the necessity for additional defects (86).

Both components of the heterodimeric transcription factor CBF, the Core Binding Factor complex (CBF), are disrupted in recurrent chromosomal translocations observed in AML. *AML1* (*RUNX1, CBFA2*), encodes the α-unit of CBF and is fused to *ETO* in the t(8;21) translocation. The β-unit of CBF is encoded by the *CBFβ* gene, which is involved in the inv(16) abnormality. The α-unit of CBF binds to DNA while the β-unit has a stabilizing role and does not have direct DNA contact (87). CBF recognizes a core DNA sequence TGT/cGGT, which is present in regulatory elements of several cellular promoters and enhancers of various hematopoietic-cell specific genes (88, 89). It has been noted that apart from the core binding sequence, adjacent binding sites for lineage-specific transcription factors, such as CEBPα and ETS family members, are also important and may direct transcription of lineage-specific genes (90, 91). These observations support the hypothesis that CBF may function as a transcriptional organizer that recruits specific factors into a complex that affects lineage-specific transcription (92). ETO, which is fused to AML1 as a result of t(8;21), is normally part of a complex, including histone deacetylase proteins, that mediates transcriptional repression (93, 94). The fusion protein AML1-ETO is capable of recruiting histone deacetylase proteins, resulting in the transcriptional repression of CBF targets such as interleukin 3 and granulocyte-macrophage colony stimulating factor (95). The fusion protein CBFβ-MYH11 can interfere with CBF DNA binding by sequestering the CBFα subunit, although the CBFβ-MYH11-CBFα complex retains the ability to bind DNA (96, 97). It is also possible that the fusion exerts its effect via local interference in transcriptional processes (96, 98). However, t(8;21) and inv(16)/t(16;16), respectively, do not cause frank leukemia, indicating that in these leukemias other abnormalities must be present (99-104).

Disruption of transcription factors is commonly seen in AML. Mixed lineage leukemia gene (MLL), the gene implicated in cytogenetical abnormalities involving 11q23, is a regulator of homeobox (HOX) genes (105, 106). The transcription factor CEBPα, that is mutated in approximately 8% of AML patients, is essential for granulocytic differentiation (107). EVI1 is a DNA binding protein and is localized in the nucleus (108). *EVI1* is involved in 3q26 abnormalities. As a result of juxtaposition of *EVI1* to the regulatory elements of the ribophorin 1 gene, *EVI1* is overexpressed. *EVI1* may also be overexpressed in cases lacking 3q26 abnormalities (109). Ectopic expression of *EVI1* in immature hematopoietic cells in vitro interferes with erythroid and

granulocytic development (110, 111).
Internal tandem duplications and tyrosine kinase domain mutations are frequently seen in *FLT3*. These lead to constitutively active receptors that activate proteins involved in signaling pathways, such as STAT5A and PI3K (112, 113). Mutations in *FLT3*, *NRAS* and *KRAS* genes are considered to be secondary mutations. They do not appear to have a role in the initiation of the neoplastic transformation, but are thought to be required for progression (114). RAS proteins transduce signals from the extracellular environment (typically initiated by receptor tyrosine kinases) to the nucleus via downstream mediator proteins such as mitogen-activated protein kinases and phosphatidylinositol 3-kinase. Mutations in *RAS* lead to impaired guanosine tri-phosphate hydrolysis and increased signal transduction and are frequently involved in human cancer. Mutations in nucleophosmin were recently identified and are found in approximately 35% of AML patients (115). Nucleophosmin is predominantly localized in the nucleus and is thought to act as a molecular chaperone, regulating the transport and assembly of pre-ribosomal proteins.
As mentioned earlier, mice expressing AML-related fusion genes such as *AML1-ETO, CBFβ-MYH11* or *PML-RARα* only develop AML at very low frequency and with long latency, indicating the necessity of secondary events (101, 116, 117). Retroviral insertional mutagenesis offers a powerful experimental strategy to identify genes involved in cancer (118, 119). Due to integration of a virus in the genome, genes are disrupted, leading to increased or decreased expression, or disrupted gene products. When the virus integrates in a critical gene or in the regulatory sequences of a critical gene, this may cause cancer. Identification of common sites of viral insertion is indicative of the presence of genes involved in leukemogenesis. In recent years, retroviral insertional mutagenesis has greatly benefited from the sequencing of the mouse genome and the development of fast polymerase chain reaction based strategies and (semi-)high throughput sequence protocols. This method has also been applied in leukemia as an approach to identifying leukemia genes and has resulted in lists of candidate proto-oncogenes (120, 121).

1.3 High-throughput assessment of RNA expression patterns

1.3.1 History of cDNA microarrays

With several publications, Mark Schena and Patrick O. Brown of the Department of Biochemistry at the Stanford University were amongst the first investigators who reported a new tool of expression monitoring: microarrays. (122-124). Base-pairing (i.e., A-T and G-C for DNA; A-U and G-C for RNA) or complementary hybridization of nucleotide sequences is the underlying principle for microarrays. Transcriptomics, the high-throughput measurement of RNA expression using microarrays, has greatly benefited from the deciphering of the human genome sequence (125, 126). On the first microarray, 1.0 kilobase (kb) cDNA (DNA complementary to RNA) sequences of 48 Arabidopsis expressed sequence tags were attached to a glass microscope slide. From total Arabidopsis mRNA, fluorescently labeled probes were prepared and hybridized to an array at high stringency. Due to this fluorescent label, each expressed sequence tag spot could be scanned using

a laser scanner and each intensity corresponded to the concentration of specific mRNA in the hybridization-mix (123). Later, this system was refined such that two hybridization-mixtures competitively bound to the same array. This facilitated comparison of mRNAs concentrations in two mixtures (127). These arrays are known as dual-channel microarrays. The two mixtures carry different fluorescent labels (Cy-3 and Cy-5) that are detected at different wavelengths. From that point, dual-channel array printing facilities were introduced at scientific institutes and microarrays were rapidly incorporated into scientific research. At the end of 1998, 35 publications reported on microarray-technology. At the end of 2000, this number had exponentially increased to 413 publications and to date, more than 12.000 articles concerned with microarrays have been published. However, the quality and content of the 'home-made' arrays greatly differed. To be able to compare results of different experiments in different institutions, standardization is needed.
Affymetrix Inc has developed a second type of microarrays, i.e., single-channel microarrays. The Affymetrix platform interrogates RNA expression using much shorter oligonucleotides (20-80 bp). It is produced using advanced photo-lithography instead of high-speed robotics (an overview of this process is described in section 1.3.3) (128). An important feature of the technique is that one single biological sample is hybridized to the array, which gives a quantitative measurement of mRNA concentrations. Other techniques for manufacturing microarrays have been developed, e.g. inkjet printing of oligonucleotides (Agilent) and piezoelectric dispensic robots which couple oligonucleotides to a glass slide covered with a polymer film (Motorola/Amersham Codelink™) (129, 130). Affymetrix GeneChips® have become a particularly frequently used platform, and currently furnish a standard in transcriptomics research.

1.3.2 History of oligonucleotide arrays

In 1991, Fodor and colleagues published an article about the light-directed synthesis of a 1024-peptide array (131). In subsequent years, this technique was optimized and focus shifted towards oligonucleotide arrays. In 1994, a study was published on the use of oligonucleotide arrays for the use of rapid DNA sequence analysis (132, 133). Although DNA sequence analysis using array technology has also proven to be feasible, this technique initially focused on the analysis of mRNA expression (128, 134). Currently, DNA sequence analysis is increasingly applied.
Golub et al. made use of Affymetrix GeneChips to classify AML and ALL samples, while Winzeler et al functionally characterized the Saccharomyces cerevisiae genome (135, 136). Both articles are now recognized as landmark studies, with currently 3254 and 902 citations, respectively.
The first commercially available oligonucleotide array was introduced in 1998 and consisted of four different arrays and represented 8000 genes (137). Due to shrinkage of feature size and more accurate data present in the sequence databases, allowing decreasing the number of probes per probe set, the number of genes present on subsequent platforms increased. Currently available oligonucleotide arrays contain approximately 55.000 probe sets, representing approximately 19.000 genes.

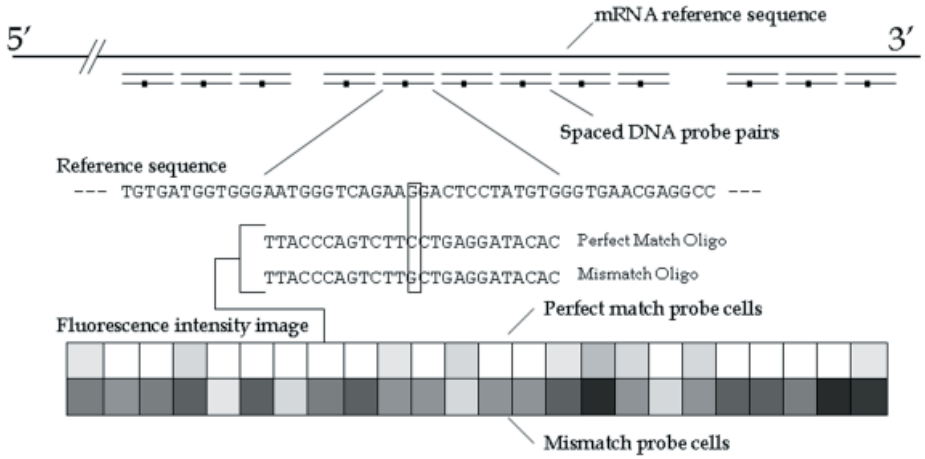


Figure 3. Overview of Affymetrix GeneChip design. *11 To 15 Perfect Match and Mismatch probes are combined in one probe set, representing one gene. Intensity values of a probe set are summarized into an expressionvalue.*

1.3.3 Synthesis and design of oligonucleotide arrays

Affymetrix oligonucleotides arrays, i.e., silicium slides of 1.28 cm by 1.28 cm, contain 25-mer oligonucleotides or probes. They are organized around probe sets and each probe set measures expression of a single gene. A probe set consists of 11 to 15 probes which are exactly complementary to cDNA (Perfect Match probes, PM) and a similar number of probes in which the middle, or 13th, nucleotide has been altered (MisMatch probes, MM) (Figure 3). All probes have been designed in the anti-sense direction. The group of probes corresponding to a given gene generally spans a region of about 600 bases at 3'-end of the start site of the gene, known as the target sequence. Most genes are represented by multiple probe sets, up to 16 probe sets for one gene, allowing some redundancy in cases of genes for which different sequences existed in different databases. Moreover, potential splice variants can be interrogated this way. The full-length genes are selected from different databases. For the HGU133 set, sequence clusters were abstracted from UniGene build 133 (April 2001), which were refined by comparison to other GenBank, RefSeq and dbEST.

Arrays are synthesized in a photolithographic process, which combines techniques from semiconductor fabrication, solid phase and combinatorial chemistry and robotics. Probes on the array are built by transmitting light to specific locations. Washing the array surface using specific nucleotide mixes sequentially generates the oligonucleotide requested on all positions. In earlier arrays, probes of the same probe set were synthesized at the same physical location. To prevent local biases, probes are currently randomly divided over the chip.

To analyze mRNA expression, RNA is isolated from the sample source. Subsequent steps are synthesis of cDNA of the RNA source, generation of biotin-labeled

antisense mRNA, fragmentation of the resulting cRNA and hybridization of the cRNA fragments on the array. After hybridization, the array is stained with a fluorescent molecule (streptavidin-phycoerythrin) that binds to the biotin-cRNA. Scanning the chip with a confocal laser results in light emission, which is converted into intensity signals.

1.3.4 Preprocessing intensity data: expression measurement

The first step in the analysis process of oligonucleotide array data is to construct expression levels from intensity values. Multiple perfect match and mismatch probes measure expression for each gene. An expression level is obtained by arithmetically combining all probes in a probe set. Several non-biological factors, such as stronger local hybridization or cross-hybridization of mRNA, can influence intensity. This may result in different background intensities and other methodological variation (138, 139). Therefore, the data have to be normalized before different experiments can be compared. Combining probe intensity signals can be done in several ways. Affymetrix has developed an algorithm of which the latest version was implemented in their analysis package Microarray Analysis Suite 5.0 (MAS5.0),which is called MAS5. This method normalizes intensities using a global scaling procedure and measures expression using a one-step Tukey biweight algorithm, which is defined as the anti-log of a robust average of differences between log(perfect match) and log(mismatch) (140). However, the a-specificity of the mismatch probes has been under debate and therefore, alternative methods have been developed (141). One of the first alternatives was the dChip-method, which scales the intensity data towards the median intensity in a group of arrays and then uses model-based index estimates, giving variable weight to perfect match-mismatch probe pairs of a probe set based on variance between arrays, to measure expression (142). Irizarry et al. introduced RMA (robust multi-array average), later followed by GCRMA (GC robust multi-array average) (143, 144). RMA, often preceded by quantile normalization (143, 145), applies a median polish procedure to PM intensities only in summarization. GCRMA is based on a similar model as RMA but takes into account the effect of stronger bonding of G/C pairs (146, 147). Other normalization and summarization methods have been developed, such as the variance stabilizing normalization (VSN) (148), which are less frequently applied. The choice which method is to be preferred is far from straightforward and several attempts have been done to assess and compare the effects of different pre-processing and normalization techniques (149-152).

1.3.5 Analysis and applications of gene expression profiles

Expression profiling has several biologically relevant applications. Microarrays have been used to identify cell cycle regulating genes in Saccharomyces cerevisiae and human cells by hybridizing cycle synchronized samples (153-156). A common application is to compare biological samples treated with and without a particular external factor, e.g. the comparative transcriptome analysis of murine skin tissue treated with and without ultraviolet light (157). Thus, expression signatures can be obtained of genes involved in particular cellular processes. Other examples of the utility of expression profiling include identification of genes involved in erythroid differentiation by comparing different lineages (158) or identification of target genes

of CEPBα by comparing CEBPα-induced and non-induced cell populations (159). Expression profiling of all specimens from patient cohorts can serve the objective of identifying different prognostic factors or identifying different disease subtypes. Gene expression profiling has for instance been applied to classify different types of myeloid leukemia (160), to predict outcome of central nervous tumors (161) and to predict de probability of developing metastasis in mamma tumors (162). Although model-based and patient cohort studies address different research questions, largely similar analysis methods are often applied.
As microarray research generates lots of data, specific analysis methods are required. To identify general patterns and structure in data, cluster analysis and principal component analysis are often applied (163, 164). These analyses classify data in different categories based on similarities, thereby identifying groups of related genes or samples (163, 164). When no prior knowledge is imposed on the data, the analysis is designated unsupervised. Supervised analysis refers to an analysis in which prior knowledge of the samples is taken into account. An example of this type of analysis is the identification of genes, which are differentially expressed in predefined subgroups of samples. In cancer research, this type of analysis can be used to identify genes that are potentially pathogenically relevant. Algorithms have been developed for this purpose, such as the Significance Analysis of Microarrays algorithm (165) or Bayesian methods (166), are being applied. Another type of supervised analysis is prediction analysis, or classification of samples. This can be performed to identify genes that are selectively expressed in a particular class of samples. Complementary to Significance Analysis of Microarrays, Prediction Analysis of Microarrays (167) has been developed for this purpose. Machine learning algorithms such as support vector machines, are also regularly applied (168). If a set of genes is found to be predictive for a particular class of samples, it could be applied in a clinical setting to predict the identity of subsequent samples.

1.4 Outline of this thesis

AML is a heterogeneous disease with a variable response to treatment (169). Classification of AML is currently based on a combination of different laboratory techniques. Previously, expression profiling has been shown to be able to distinguish myeloid from lympoid leukemia, and distinct subtypes within these diseases (136, 160, 170, 171).
The first goal of the work described in this thesis is to investigate whether expression profiling can be used to classify distinct AML subgroups with one comprehensive assay. Chapter 2 describes the results of an expression profiling study using the bone marrow and blood specimens of a cohort of 285 de novo AML patients.
Several statistical pre-processing methods exist to combine data obtained with oligonucleotide arrays into gene expression levels (140, 142, 143). Different pre-processing methods may lead to different expression estimates and may therefore have an influence on the outcome of both supervised and unsupervised analyses. Chapter 3 describes an assessment of the magnitude of this effect.
Recently, Falini and colleagues described the abberant cytoplasmic location of a new oncogene, i.e. nucleophosmin or *NPM1*, in approximately 35% of AML patients (115). Nucleophosmin is thought to function mainly as a molecular chaperone of proteins, facilitating the transport of ribosomal proteins through the

nuclear membrane. Mutations in *NPM1* were described to result in its abnormal cytoplasmic location. Chapter 4 describes the presence, nature, and expression characteristics and the clinical prognostic value of these mutations in a cohort of 275 AML patients.
A variety of genes are differentially expressed in different subtypes of AML. Differentially expressed genes could be involved in the pathogenesis of AML, but could also relate to phenotypic variations of different leukemias, e.g. related to the maturation status. Chapter 5 specifically describes a methodology to distinguish pathogenetically relevant genes among large sets of differentially expressed genes in clinical AML by including results from retrovirally induced mouse leukemias.
Acute interpretation of data obtained by unsupervised analysis of large scale expression profiling studies is currently frequently performed by visually combining sample-gene heatmaps and sample characteristics. In Chapter 6, we present an implementation of an integration method for such visualizations.
At the end, the results described in this thesis are discussed (Chapter 7).

References

1. Moore MA, Metcalf D. Ontogeny of the haemopoietic system: yolk sac origin of in vivo and in vitro colony forming cells in the developing mouse embryo. Br J Haematol 1970;18(3):279-96.
2. Dieterlen-Lievre F. On the origin of haemopoietic stem cells in the avian embryo: an experimental approach. J Embryol Exp Morphol 1975;33(3):607-19.
3. Turpen JB, Knudson CM, Hoefen PS. The early ontogeny of hematopoietic cells studied by grafting cytogenetically labeled tissue anlagen: localization of a prospective stem cell compartment. Dev Biol 1981;85(1):99-112.
4. Medvinsky AL, Samoylina NL, Muller AM, Dzierzak EA. An early pre-liver intraembryonic source of CFU-S in the developing mouse. Nature 1993;364(6432):64-7.
5. Medvinsky AL, Gan OI, Semenova ML, Samoylina NL. Development of day-8 colony-forming unit-spleen hematopoietic progenitors during early murine embryogenesis: spatial and temporal mapping. Blood 1996;87(2):557-66.
6. Medvinsky A, Dzierzak E. Definitive hematopoiesis is autonomously initiated by the AGM region. Cell 1996;86(6):897-906.
7. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 2004;20(3):323-31.
8. Mikkola HK, Gekas C, Orkin SH, Dieterlen-Lievre F. Placenta as a site for hematopoietic stem cell development. Exp Hematol 2005;33(9):1048-54.
9. Johnson GR, Moore MA. Role of stem cell migration in initiation of mouse foetal liver haemopoiesis. Nature 1975;258(5537):726-8.
10. Jordan CT, Lemischka IR. Clonal and systemic analysis of long-term hematopoiesis in the mouse. Genes Dev 1990;4(2):220-32.
11. Abramson S, Miller RG, Phillips RA. The identification in adult bone marrow of pluripotent and restricted stem cells of the myeloid and lymphoid systems. J Exp Med 1977;145(6):1567-9.
12. Dick JE, Magli MC, Huszar D, Phillips RA, Bernstein A. Introduction of a selectable gene into primitive stem cells capable of long-term reconstitution of the hemopoietic system of W/Wv mice. Cell 1985;42(1):71-9.
13. Keller G, Paige C, Gilboa E, Wagner EF. Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. Nature 1985;318(6042):149-54.
14. Szilvassy SJ, Cory S. Efficient retroviral gene transfer to purified long-term repopulating hematopoietic stem cells. Blood 1994;84(1):74-83.
15. Szilvassy SJ. The biology of hematopoietic stem cells. Arch Med Res 2003;34(6):446-60.

16. Jordan CT, Lemischka IR. Clonal and systematic analysis of long-term hematopoiesis in the mouse. Genes Dev 1990;4:220-32.
17. Zandstra PW, Lauffenburger DA, Eaves CJ. A ligand-receptor signaling threshold model of stem cell differentiation control: a biologically conserved mechanism applicable to hematopoiesis. Blood 2000;96(4):1215-22.
18. Gordon MY. Hemopoietic growth factors and receptors: bound and free. Cancer Cells 1991;3(4):127-33.
19. Keller G, Snodgrass R. Life span of multipotential hematopoietic stem cells in vivo. J Exp Med 1990;171(5):1407-18.
20. Capel B, Hawley RG, Mintz B. Long- and short-lived murine hematopoietic stem cell clones individually identified with retroviral integration markers. Blood 1990;75(12):2267-70.
21. Krause DS, Theise ND, Collector MI, et al. Multi-organ, multi-lineage engraftment by a single bone marrow-derived stem cell. Cell 2001;105(3):369-77.
22. Osawa M, Hanada K, Hamada H, Nakauchi H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. Science 1996;273(5272):242-5.
23. Morrison SJ, Hemmati HD, Wandycz AM, Weissman IL. The purification and characterization of fetal liver hematopoietic stem cells. Proc Natl Acad Sci U S A 1995;92(22):10302-6.
24. Spangrude GJ, Heimfeld S, Weissman IL. Purification and characterization of mouse hematopoietic stem cells. Science 1988;241(4861):58-62.
25. Miles C, Sanchez MJ, Sinclair A, Dzierzak E. Expression of the Ly-6E.1 (Sca-1) transgene in adult hematopoietic stem cells and the developing mouse embryo. Development 1997;124(2):537-47.
26. Sanchez MJ, Holmes A, Miles C, Dzierzak E. Characterization of the first definitive hematopoietic stem cells in the AGM and liver of the mouse embryo. Immunity 1996;5(6):513-25.
27. Lowenberg B, Downing JR, Burnett A. Acute myeloid leukemia. N Engl J Med 1999;341(14):1051-62.
28. Bol P. Acute myeloide leukemie. Nederlands Tijdschrift voor Tandheelkunde 2002;109(11):463-4.
29. KWF. 2000. (Accessed at http://www.kwf.nl.)
30. Bennett JM, Catovsky D, Daniel MT, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. Br J Haematol 1976;33(4):451-8.
31. Bennett JM, Catovsky D, Daniel MT, et al. A variant form of hypergranular promyelocytic leukaemia (M3). Br J Haematol 1980;44(1):169-70.
32. Bennett JM, Catovsky D, Daniel MT, et al. Criteria for the diagnosis of acute leukemia of megakaryocyte lineage (M7). A report of the French-American-British Cooperative Group. Ann Intern Med 1985;103(3):460-2.
33. 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 1985;103(4):620-5.
34. Jaffe ES, Harris NL, Vardiman JW, Stein N, eds. World Health Organization Classification of Tumours. Pathology and Genetic of Tumours of Haematopoietic and Lymphoid Tissues. Lyon: IARC Press; 2001.
35. Vardiman JW, Harris NL, Brunning RD. The World Health Organization (WHO) classification of the myeloid neoplasms. Blood 2002;100(7):2292-302.
36. Lowenberg B. Prognostic factors in acute myeloid leukaemia. Best Pract Res Clin Haematol 2001;14(1):65-75.
37. Keating MJ, Smith TL, Kantarjian H, et al. Cytogenetic pattern in acute myelogenous leukemia: a major reproducible determinant of outcome. Leukemia 1988;2(7):403-12.
38. Samuels BL, Larson RA, Le Beau MM, et al. Specific chromosomal abnormalities in acute nonlymphocytic leukemia correlate with drug susceptibility in vivo. Leukemia 1988;2(2):79-83.
39. Berger R, Bernheim A, Ochoa-Noguera ME, et al. Prognostic significance of chromosomal abnormalities in acute nonlymphocytic leukemia: a study of 343 patients. Cancer Genet

Cytogenet 1987;28(2):293-9.
40. Marosi C, Koller U, Koller-Weber E, et al. Prognostic impact of karyotype and immunologic phenotype in 125 adult patients with de novo AML. Cancer Genet Cytogenet 1992;61(1):14-25.
41. Arthur DC, Berger R, Golomb HM, et al. The clinical significance of karyotype in acute myelogenous leukemia. Cancer Genet Cytogenet 1989;40(2):203-16.
42. Fenaux P, Preudhomme C, Lai JL, Morel P, Beuscart R, Bauters F. Cytogenetics and their prognostic value in de novo acute myeloid leukaemia: a report on 283 cases. Br J Haematol 1989;73(1):61-7.
43. Grimwade D, Walker H, Oliver F, et al. The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. Blood 1998;92(7):2322-33.
44. Swansbury GJ, Lawler SD, Alimena G, et al. Long-term survival in acute myelogenous leukemia: a second follow-up of the Fourth International Workshop on Chromosomes in Leukemia. Cancer Genet Cytogenet 1994;73(1):1-7.
45. Dastugue N, Payen C, Lafage-Pochitaloff M, et al. Prognostic significance of karyotype in de novo adult acute myeloid leukemia. The BGMT group. Leukemia 1995;9(9):1491-8.
46. Bloomfield CD, Shuma C, Regal L, et al. Long-term survival of patients with acute myeloid leukemia: a third follow-up of the Fourth International Workshop on Chromosomes in Leukemia. Cancer 1997;80(11 Suppl):2191-8.
47. Cassileth PA, Harrington DP, Appelbaum FR, et al. Chemotherapy compared with autologous or
allogeneic bone marrow transplantation in the management of acute myeloid leukemia in first remission. N Engl J Med 1998;339(23):1649-56.
48. Wheatley K, Burnett AK, Goldstone AH, et al. A simple, robust, validated and highly predictive index for the determination of risk-directed therapy in acute myeloid leukaemia derived from the MRC AML 10 trial. United Kingdom Medical Research Council's Adult and Childhood Leukaemia Working Parties. Br J Haematol 1999;107(1):69-79.
49. Nakao M, Yokota S, Iwai T, et al. Internal tandem duplication of the flt3 gene found in acute myeloid leukemia. Leukemia 1996;10(12):1911-8.
50. 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 2002;99(12):4326-35.
51. 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.
52. 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 2002;100(8):2717-23.
53. Rombouts WJ, Blokland I, Lowenberg B, Ploemacher RE. Biological characteristics and prognosis of adult acute myeloid leukemia with internal tandem duplications in the Flt3 gene. Leukemia 2000;14(4):675-83.
54. Kiyoi H, Naoe T, Nakano Y, et al. Prognostic implication of FLT3 and N-RAS gene mutations in acute myeloid leukemia. Blood 1999;93(9):3074-80.
55. Barjesteh van Waalwijk van Doorn-Khosrovani S, Erpelinck C, van Putten WL, et al. High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. Blood 2003;101(3):837-45.
56. Karakas T, Miething CC, Maurer U, et al. The coexpression of the apoptosis-related genes bcl-2 and wt1 in predicting survival in adult acute myeloid leukemia. Leukemia 2002;16(5):846-54.
57. Nakano Y, Naoe T, Kiyoi H, et al. Prognostic value of p53 gene mutations and the product expression in de novo acute myeloid leukemia. Eur J Haematol 2000;65(1):23-31.
58. Lowenberg B, Touw IP. Hematopoietic growth factors and their receptors in acute leukemia. Blood 1993;81(2):281-92.
59. Lowenberg B, van Putten WL, Touw IP, Delwel R, Santini V. Autonomous proliferation of leukemic cells in vitro as a determinant of prognosis in adult acute myeloid leukemia. N

Engl J Med 1993;328(9):614-9.
60. Kosary C, Ries L, Miller B, Hankey B, Edwards B. SEER cancer statistics review, 1973-1992: tables and graphs.: National Cancer Institute; 1995. Report No.: 96-2789.
61. Fenaux P, Chastang C, Chevret S, et al. A randomized comparison of all transretinoic acid (ATRA) followed by chemotherapy and ATRA plus chemotherapy and the role of maintenance therapy in newly diagnosed acute promyelocytic leukemia. The European APL Group. Blood 1999;94(4):1192- 200.
62. Druker BJ, Talpaz M, Resta DJ, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 2001;344(14):1031-7.
63. Gorre ME, Sawyers CL. Molecular mechanisms of resistance to STI571 in chronic myeloid leukemia. Curr Opin Hematol 2002;9(4):303-7.
64. Stone RM, DeAngelo DJ, Klimek V, et al. Patients with acute myeloid leukemia and an activating mutation in FLT3 respond to a small-molecule FLT3 tyrosine kinase inhibitor, PKC412. Blood 2005;105(1):54-60.
65. Karp JE, Lancet JE, Kaufmann SH, et al. Clinical and biologic activity of the farnesyltransferase inhibitor R115777 in adults with refractory and relapsed acute leukemias: a phase 1 clinical-laboratory correlative trial. Blood 2001;97(11):3361-9.
66. Kelly LM, Liu Q, Kutok JL, Williams IR, Boulton CL, Gilliland DG. FLT3 internal tandem duplication mutations associated with human acute myeloid leukemias induce myeloproliferative disease in a murine bone marrow transplant model. Blood 2002;99(1):310-8.
67. Wilbanks AM, Mahajan S, Frank DA, Druker BJ, Gilliland DG, Carroll M. TEL/PDGFbetaR fusion protein activates STAT1 and STAT5: a common mechanism for transformation by tyrosine kinase fusion proteins. Exp Hematol 2000;28(5):584-93.
68. Schwaller J, Frantsve J, Aster J, et al. Transformation of hematopoietic cell lines to growth-factor independence and induction of a fatal myelo- and lymphoproliferative disease in mice by retrovirally transduced TEL/JAK2 fusion genes. Embo J 1998;17(18):5321-33.
69. Kundu M, Liu PP. Function of the inv(16) fusion gene CBFB-MYH11. Curr Opin Hematol 2001;8(4):201-5.
70. Licht JD. AML1 and the AML1-ETO fusion protein in the pathogenesis of t(8;21) AML. Oncogene 2001;20(40):5660-79.
71. Melnick A, Licht JD. Deconstructing a disease: RARalpha, its fusion partners, and their roles in the pathogenesis of acute promyelocytic leukemia. Blood 1999;93(10):3167-215.
72. Shigesada K, van de Sluis B, Liu PP. Mechanism of leukemogenesis by the inv(16) chimeric gene CBFB/PEBP2B-MHY11. Oncogene 2004;23(24):4297-307.
73. Cohen N, Sharma M, Kentsis A, Perez JM, Strudwick S, Borden KL. PML RING suppresses oncogenic transformation by reducing the affinity of eIF4E for mRNA. Embo J 2001;20(16):4547-59.
74. Mu ZM, Chin KV, Liu JH, Lozano G, Chang KS. PML, a growth suppressor disrupted in acute promyelocytic leukemia. Mol Cell Biol 1994;14(10):6858-67.
75. Fogal V, Gostissa M, Sandy P, et al. Regulation of p53 activity in nuclear bodies by a specific PML isoform. Embo J 2000;19(22):6185-95.
76. Guo A, Salomoni P, Luo J, et al. The function of PML in p53-dependent apoptosis. Nat Cell Biol 2000;2(10):730-6.
77. Torii S, Egan DA, Evans RA, Reed JC. Human Daxx regulates Fas-induced apoptosis from nuclear PML oncogenic domains (PODs). Embo J 1999;18(21):6037-49.
78. Zhong S, Salomoni P, Ronchetti S, Guo A, Ruggero D, Pandolfi PP. Promyelocytic leukemia protein (PML) and Daxx participate in a novel nuclear pathway for apoptosis. J Exp Med 2000;191(4):631-40.
79. Alcalay M, Tomassoni L, Colombo E, et al. The promyelocytic leukemia gene product (PML) forms stable complexes with the retinoblastoma protein. Mol Cell Biol 1998;18(2):1084-93.
80. Rosenbauer F, Wagner K, Kutok JL, et al. Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, PU.1. Nat Genet 2004;36(6):624-30.
81. Vallian S, Chin KV, Chang KS. The promyelocytic leukemia protein interacts with Sp1 and inhibits its transactivation of the epidermal growth factor receptor promoter. Mol Cell Biol 1998;18(12):7147-56.
82. Chambon P. A decade of molecular biology of retinoic acid receptors. Faseb J 1996;10(9):940-

54.
83. Nervi C, Poindexter EC, Grignani F, et al. Characterization of the PML-RAR alpha chimeric product of the acute promyelocytic leukemia-specific t(15;17) translocation. Cancer Res 1992;52(13):3687-92.
84. Grunstein M. Histone acetylation in chromatin structure and transcription. Nature 1997;389(6649):349-52.
85. Lin RJ, Nagy L, Inoue S, Shao W, Miller WH, Jr., Evans RM. Role of the histone deacetylase complex in acute promyelocytic leukaemia. Nature 1998;391(6669):811-4.
86. Mistry AR, Pedersen EW, Solomon E, Grimwade D. The molecular pathogenesis of acute promyelocytic leukaemia: implications for the clinical management of the disease. Blood Rev 2003;17(2):71-97.
87. 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 1993;13(10):6336-45.
88. Takahashi A, Satake M, Yamaguchi-Iwai Y, et al. Positive and negative regulation of granulocyte-macrophage colony-stimulating factor promoter activity by AML1-related transcription factor, PEBP2. Blood 1995;86(2):607-16.
89. Zhang DE, Fujioka K, Hetherington CJ, et al. Identification of a region which directs the monocytic activity of the colony-stimulating factor 1 (macrophage colony-stimulating factor) receptor promoter and binds PEBP2/CBF (AML1). Mol Cell Biol 1994;14(12):8085-95.
90. Wotton D, Ghysdael J, Wang S, Speck NA, Owen MJ. Cooperative binding of Ets-1 and core binding factor to DNA. Mol Cell Biol 1994;14(1):840-50.
91. Zhang DE, Hohaus S, Voso MT, et al. Function of PU.1 (Spi-1), C/EBP, and AML1 in early myelopoiesis: regulation of multiple myeloid CSF receptor promoters. Curr Top Microbiol Immunol 1996;211:137-47.
92. Carey M. The enhanceosome and transcriptional synergy. Cell 1998;92(1):5-8.
93. Amann JM, Nip J, Strom DK, et al. ETO, a target of t(8;21) in acute leukemia, makes distinct contacts with multiple histone deacetylases and binds mSin3A through its oligomerization domain. Mol Cell Biol 2001;21(19):6470-83.
94. Wang J, Hoshino T, Redner RL, Kajigaya S, Liu JM. ETO, fusion partner in t(8;21) acute myeloid leukemia, represses transcription by interaction with the human N-CoR/mSin3/HDAC1 complex. Proc Natl Acad Sci U S A 1998;95(18):10860-5.
95. Hart SM, Foroni L. Core binding factor genes and human leukemia. Haematologica 2002;87(12):1307-23.
96. Cao W, Britos-Bray M, Claxton DF, et al. CBF beta-SMMHC, expressed in M4Eo AML, reduced CBF DNA-binding and inhibited the G1 to S cell cycle transition at the restriction point in myeloid and lymphoid cells. Oncogene 1997;15(11):1315-27.
97. Liu PP, Wijmenga C, Hajra A, et al. Identification of the chimeric protein product of the CBFB-MYH11 fusion gene in inv(16) leukemia cells. Genes Chromosomes Cancer 1996;16(2):77-87.
98. Liu P, Tarle SA, Hajra A, et al. Fusion between transcription factor CBF beta/PEBP2 beta and a myosin heavy chain in acute myeloid leukemia. Science 1993;261(5124):1041-4.
99. Castilla LH, Garrett L, Adya N, et al. The fusion gene Cbfb-MYH11 blocks myeloid differentiation and predisposes mice to acute myelomonocytic leukaemia. Nat Genet 1999;23(2):144-6.
100. Castilla LH, Perrat P, Martinez NJ, et al. Identification of genes that synergize with Cbfb-MYH11 in the pathogenesis of acute myeloid leukemia. Proc Natl Acad Sci U S A 2004;101(14):4924-9.
101. Higuchi M, O'Brien D, Kumaravelu P, Lenny N, Yeoh EJ, Downing JR. Expression of a conditional AML1-ETO oncogene bypasses embryonic lethality and establishes a murine model of human t(8;21) acute myeloid leukemia. Cancer Cell 2002;1(1):63-74.
102. Okuda T, Cai Z, Yang S, et al. Expression of a knocked-in AML1-ETO leukemia gene inhibits the establishment of normal definitive hematopoiesis and directly generates dysplastic hematopoietic progenitors. Blood 1998;91(9):3134-43.
103. Yergeau DA, Hetherington CJ, Wang Q, et al. Embryonic lethality and impairment of haematopoiesis in mice heterozygous for an AML1-ETO fusion gene. Nat Genet

1997;15(3):303-6.
104. Yuan Y, Zhou L, Miyamoto T, et al. AML1-ETO expression is directly involved in the development of acute myeloid leukemia in the presence of additional mutations. Proc Natl Acad Sci U S A 2001;98(18):10398-403.
105. Yu BD, Hanson RD, Hess JL, Horning SE, Korsmeyer SJ. MLL, a mammalian trithorax-group gene, functions as a transcriptional maintenance factor in morphogenesis. Proc Natl Acad Sci U S A 1998;95(18):10632-6.
106. Yu BD, Hess JL, Horning SE, Brown GA, Korsmeyer SJ. Altered Hox expression and segmental identity in Mll-mutant mice. Nature 1995;378(6556):505-8.
107. Radomska HS, Huettner CS, Zhang P, Cheng T, Scadden DT, Tenen DG. CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. Mol Cell Biol 1998;18(7):4301-14.
108. Jolkowska J, Witt M. The EVI-1 gene--its role in pathogenesis of human leukemias. Leuk Res 2000;24(7):553-8.
109. Suzukawa K, Parganas E, Gajjar A, et al. Identification of a breakpoint cluster region 3' of the ribophorin I gene at 3q21 associated with the transcriptional activation of the EVI1 gene in acute myelogenous leukemias with inv(3)(q21q26). Blood 1994;84(8):2681-8.
110. Kreider BL, Orkin SH, Ihle JN. Loss of erythropoietin responsiveness in erythroid progenitors due to expression of the Evi-1 myeloid-transforming gene. Proc Natl Acad Sci U S A 1993;90(14):6454-8.
111. Morishita K, Parganas E, Matsugi T, Ihle JN. Expression of the Evi-1 zinc finger gene in 32Dc13 myeloid cells blocks granulocytic differentiation in response to granulocyte colony-stimulating factor. Mol Cell Biol 1992;12(1):183-9.
112. Zhang S, Fukuda S, Lee Y, et al. Essential role of signal transducer and activator of transcription (Stat)5a but not Stat5b for Flt3-dependent signaling. J Exp Med 2000;192(5):719-28.
113. Zhang S, Broxmeyer HE. Flt3 ligand induces tyrosine phosphorylation of gab1 and gab2 and their association with shp-2, grb2, and PI3 kinase. Biochem Biophys Res Commun 2000;277(1):195-9.
114. Moore MA. Converging pathways in leukemogenesis and stem cell self-renewal. Exp Hematol 2005;33(7):719-37.
115. Falini B, Mecucci C, Tiacci E, et al. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. N Engl J Med 2005;352(3):254-66.
116. Johnson JJ, Chen W, Hudson W, et al. Prenatal and postnatal myeloid cells demonstrate stepwise progression in the pathogenesis of MLL fusion gene leukemia. Blood 2003;101(8):3229-35.
117. Kundu M, Chen A, Anderson S, et al. Role of Cbfb in hematopoiesis and perturbations resulting from expression of the leukemogenic fusion gene Cbfb-MYH11. Blood 2002;100(7):2449-56.
118. Jonkers J, Berns A. Retroviral insertional mutagenesis as a strategy to identify cancer genes. Biochim Biophys Acta 1996;1287(1):29-57.
119. Suzuki T, Shen H, Akagi K, et al. New genes involved in cancer identified by retroviral tagging. Nat Genet 2002;32(1):166-74.
120. Erkeland SJ, Valkhof M, Heijmans-Antonissen C, et al. Large-scale identification of disease genes involved in acute myeloid leukemia. J Virol 2004;78(4):1971-80.
121. Li J, Shen H, Himmel KL, et al. Leukaemia disease genes: large-scale cloning and pathway predictions. Nat Genet 1999;23(3):348-53.
122. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A 1996;93(20):10614-9.
123. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270(5235):467-70.
124. DeRisi J, Penland L, Brown PO, et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14(4):457-60.
125. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.
126. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science

2001;291(5507):1304-51.
127. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6(7):639-45.
128. Lockhart DJ, Dong H, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14(13):1675-80.
129. Ramakrishnan R, Dorris D, Lublinsky A, et al. An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. Nucleic Acids Res 2002;30(7):e30.
130. Hughes TR, Mao M, Jones AR, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat Biotechnol 2001;19(4):342-7.
131. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251(4995):767-73.
132. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. Nature 1993;364(6437):555-6.
133. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci U S A 1994;91(11):5022-6.
134. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274(5287):610-4.
135. Winzeler EA, Shoemaker DD, Astromoff A, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 1999;285(5429):901-6.
136. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531-7.
137. Affymetrix I. GeneChip Expression Platform: Comparison, Evolution, and Performance. Santa Clara; 2003.
138. Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem Suppl 2001;Suppl 37:120-5.
139. Schadt EE, Li C, Su C, Wong WH. Analyzing high-density oligonucleotide gene expression array data. J Cell Biochem 2000;80(2):192-202.
140. Affymetrix. Microarray Suite User Guide; 2001.
141. Forman JE, Walton ID, Stern D, Rava RP, Trulson MO. Thermodynamics of duplex formation and mismatch discrimination on photolithograpically synthesized oligonucleotide arrays. ACS Symp Ser 1998;682:206-28.
142. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 2001;98(1):31-6.
143. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31(4):e15.
144. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model based background adjustment for oligonucleotide expression arrays. Journal of American Statistical Association 2004;99(468):909-17.
145. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-93.
146. Naef F, Magnasco MO. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Phys Rev E Stat Nonlin Soft Matter Phys 2003;68(1 Pt 1):011906.
147. Wu Z, Irizarry RA, Gentleman R, Murillo F, Spencer F. A model based background adjustment for oligonucleotide expression arrays. Technical report. Baltimore: John Hopkins University; 2004.
148. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;18 Suppl 1:S96-104.
149. Rajagopalan D. A comparison of statistical methods for analysis of high density oligonucleotide array data. Bioinformatics 2003;19(12):1469-76.
150. Hoffmann R, Seidl T, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol

2002;3(7):RESEARCH0033.
151. Freudenberg J, Boriss H, Hasenclever D. Comparison of preprocessing procedures for oligo-nucleotide micro-arrays by parametric bootstrap simulation of spike-in experiments. Methods Inf Med 2004;43(5):434-8.
152. Shedden K, Chen W, Kuick R, et al. Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. BMC Bioinformatics 2005;6(1):26.
153. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998;9(12):3273-97.
154. Shedden K, Cooper S. Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. Proc Natl Acad Sci U S A 2002;99(7):4379-84.
155. Shedden K, Cooper S. Analysis of cell-cycle gene expression in Saccharomyces cerevisiae using microarrays and multiple synchronization methods. Nucleic Acids Res 2002;30(13):2920-9.
156. Wolfsberg TG, Gabrielian AE, Campbell MJ, Cho RJ, Spouge JL, Landsman D. Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. Genome Res 1999;9(8):775-92.
157. Garinis GA, Mitchell JR, Moorhouse MJ, et al. Transcriptome analysis reveals cyclobutane pyrimidine dimers as a major source of UV-induced DNA breaks. Embo J 2005;24(22):3952-62.
158. Heo HS, Kim JH, Lee YJ, Kim SH, Cho YS, Kim CG. Microarray profiling of genes differentially expressed during erythroid differentiation of murine erythroleukemia cells. Mol Cells 2005;20(1):57-68.
159. Gery S, Gombart AF, Yi WS, Koeffler C, Hofmann WK, Koeffler HP. Transcription profiling of C/EBP targets identifies Per2 as a gene implicated in myeloid leukemia. Blood 2005;106(8):2827-36.
160. Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30(1):41-7.
161. Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002;415(6870):436-42.
162. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530-6.
163. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863-8.
164. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000:455-66.
165. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001;98(9):5116-21.
166. Smyth GK. Lineair models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 2004;3(1).
167. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99(10):6567-72.
168. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16(10):906-14.
169. Giles FJ, Keating A, Goldstone AH, Avivi I, Willman CL, Kantarjian HM. Acute myeloid leukemia. Hematology (Am Soc Hematol Educ Program) 2002:73-110.
170. Debernardi S, Lillington DM, Chaplin T, et al. Genome-wide analysis of acute myeloid leukemia with normal karyotype reveals a unique pattern of homeobox gene expression distinct from those with translocation-mediated fusion events. Genes Chromosomes Cancer 2003;37(2):149-58.
171. Schoch C, Kohlmann A, Schnittger S, et al. Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. Proc Natl Acad

Sci U S A 2002;99(15):10008-13.

Chapter 2

Prognostically useful gene-expression profiles in acute myeloid leukemia

Peter J.M. Valk[1], Roel G.W. Verhaak[1], M. Antoinette Beijen[1], Claudia A.J. Erpelinck[1], Sahar Barjesteh van Waalwijk van Doorn - Khosrovani[1], Judith M. Boer[2], H. Berna Beverloo[3], Michael J. Moorhouse[4], Peter J. van der Spek[4], Bob Löwenberg[1], Ruud Delwel[1]

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

[2]Leiden Genome Technology Center (LGTC), Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands

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

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

New England Journal of Medicine. 2004 Apr 15;350(16):1617-28.

2.1 Abstract

2.1.1 Background

In patients with acute myeloid leukemia (AML) a combination of methods must be used to classify the disease, make therapeutic decisions, and determine the prognosis. However, this combined approach provides correct therapeutic and prognostic information in only 50 percent of cases.

2.1.2 Methods

We determined the gene-expression profiles in samples of peripheral blood or bone marrow from 285 patients with AML using Affymetrix U133A GeneChips containing approximately 13,000 unique genes or expression-signature tags. Data analyses were carried out with Omniviz, significance analysis of microarrays, and prediction analysis of microarrays software. Statistical analyses were performed to determine the prognostic significance of cases of AML with specific molecular signatures.

2.1.3 Results

Unsupervised cluster analyses identified 16 groups of patients with AML on the basis of molecular signatures. We identified the genes that defined these clusters and determined the minimal numbers of genes needed to identify prognostically important clusters with a high degree of accuracy. The clustering was driven by the presence of chromosomal lesions (e.g., t(8;21), t(15;17), and inv(16)), particular genetic mutations (*CEBPa*), and abnormal oncogene expression (*EVI1*). We identified several novel clusters, some consisting of specimens with normal karyotypes. A unique cluster with a distinctive gene-expression signature included cases of AML with a poor treatment outcome.

2.1.4 Conclusions

Gene-expression profiling allows a comprehensive classification of AML that includes previously identified genetically defined subgroups and a novel cluster with an adverse prognosis.

2.2 Introduction

Acute myeloid leukemia (AML) is not a single disease but a group of neoplasms with diverse genetic abnormalities and variable responses to treatment. Cytogenetics and molecular analyses can be used to identify subgroups of AML with different prognoses. For instance, the translocations inv(16), t(8;21), and t(15;17) herald a favorable prognosis, whereas other cytogenetic aberrations indicate poor-risk leukemia (1-5). Abnormalities involving 11q23, t(6;9), or 7(q) are defined as poor-risk markers by some groups (2,3) and as intermediate- risk markers by others (3-5). These inconsistencies and the absence of cytogenetic abnormalities in a considerable proportion of patients argue for refinement of the classification of AML. Additional reasons for extending the molecular analyses of AML are exemplified by findings regarding the gene for fms-like tyrosine kinase 3 (*FLT3*), the gene encoding ectotropic viral integration 1 site (*EVI1*), and the gene for CCAAT/

enhancer binding protein alpha (*CEBPa*). An internal tandem duplication in *FLT3*, a hematopoietic growth factor receptor, is the most common molecular abnormality in AML (6,7). The presence of such mutations in *FLT3* and elevated expression of the transcription factor EVI1 confer a poor prognosis, (6-8) whereas mutations in *CEBPa* are associated with a good outcome (9,10). Molecular classification based on DNA-expression profiling offers a powerful way of distinguishing myeloid from lymphoid cancer and subclasses within these two diseases (11-14). DNA-microarray analysis has the potential to identify distinct subgroups of AML with the use of one comprehensive assay, to classify cases that currently resist categorization by means of other methods, and to identify subgroups with favorable or unfavorable prognoses within genetically defined subclasses. The goals of this study of 285 adults with AML were to use gene expression profiles to identify established and novel subclasses of AML and otherwise unrecognized cases of poor-risk AML.

2.3 Methods

2.3.1 Patients and cell samples

Eligible patients had received a diagnosis of primary AML, which had been confirmed by means of a cytologic examination of blood and bone marrow (Table 1). All patients were treated according to the protocols of the Dutch–Belgian Hematology–Oncology Cooperative group (available at www.hovon.nl) (15-17). All subjects provided written informed consent. A total of 285 patients provided bone marrow aspirates or peripheral-blood samples at the time of diagnosis and 8 healthy control subjects provided peripheral-blood samples or bone marrow aspirates. Blasts and mononuclear cells were purified by Ficoll–Hypaque (Nygaard) centrifugation and cryopreserved. CD34+ cells from three control subjects were sorted by means of a fluorescence-activated cell sorter. The AML samples contained 80 to 100 percent blast cells after thawing, regardless of the blast count at diagnosis.

2.3.2 Isolation and quality control of RNA

After thawing, cells were washed once with Hanks'balanced-salt solution. High-quality total RNA was extracted by lysis with guanidinium thiocyanate followed by cesium chloride–gradient purification (18). RNA levels, quality, and purity were assessed with the use of the RNA 6000 Nano assay on the Agilent 2100 Bioanalyzer (Agilent). None of the samples showed RNA degradation (ratio of 28S ribosomal RNA to 18S ribosomal RNA of at least 2) or contamination by DNA.

2.3.3 Gene profiling and quality control

Samples were analyzed with the use of Affymetrix U133A GeneChips. Each gene on this chip is represented by 10 to 20 oligonucleotides, termed a "probe set." The intensity of hybridization of labeled messenger RNA (mRNA) to these sets reflects the level of expression of a particular gene. The U133A GeneChip contains 22,283 probe sets, representing approximately 13,000 genes. We used 10 µg of total RNA to prepare antisense biotinylated RNA. Single-stranded complementary DNA (cDNA) and double-stranded cDNA were synthesized according to the manufacturer's protocol (Invitrogen Life Technologies) with the use of the

Glossary

Centroid: In a self-organizing topologic map of gene expression, the centroid corresponds to the center of a cluster.

Chromosomal abnormalities

t(8;21): One of the commonest cytogenetic abnormalities in AML; produces a hybrid gene by fusing *AML1* on the long arm of chromosome 21 with *ETO* on the long arm of chromosome 8.

inv(16): Inversion of a segment of chromosome 16 that produces the *CBFβ-MYH11* fusion.

t(15;17): Reciprocal translocation of genetic material between the long arms of chromosomes 15 and 17 that produces the *PML-RARa* fusion gene, typical of acute promyelocytic leukemia.

11q23: A chromosomal region that becomes rearranged with various partner chromosomal regions in diverse forms of leukemia, involving the MLL gene.

t(6;9): A rare translocation often found in young patients and sometimes associated with basophilia.

-7(q): Loss of the long arm of chromosome 7, on monosomy 7.

French–American–British (FAB) classification: An internationally agreed-on method of classifying acute leukemia by morphologic means. There are eight subtypes, ranging from M0 (myeloblasts) to M7 (megakaryoblasts).

Gene-expression profiling: Determination of the level of expression of thousands of genes through the use of microarrays. Messenger RNA extracted from the test tissue or cells and labeled with a fluorescent dye is tested for its ability to hybridize to the spotted nucleic acids.

Microarray or GeneChip: A robotically spotted array of thousands of complementary DNAs or oligonucleotides.

Patient-clustering technique: A method of grouping patients with similar patterns of gene expression.

Pearson's correlation coefficient: A statistical measure of the strength of the relationship between variables.

Pearson's Correlation Visualization tool of Omniviz: Omniviz is a commercial multifunctional statistical package used for analysis of microarray data. It allows the visual representation of gene-expression profiles of patients in a Pearson's Correlation View.

Prediction analysis of microarrays (PAM): A statistical technique that identifies a subgroup of genes that best characterizes a predefined class.

Probe set: A group of 10 to 20 oligonucleotides; each set corresponds to one gene.

Significance analysis of microarrays (SAM): A statistical method used in microarray analyses that identifies genes that are significantly differentially expressed between groups of patients on the basis of a change in the level of gene expression relative to the standard deviation of repeated measurements.

Supervised analysis: An analysis of the results of microarray profiling that takes external factors into account.

Unsupervised analysis: An analysis of the results of microarray profiling that does not take external factors such as survival or clinical signs into account.

10-Fold cross-validation: A validation method that works as follows: the model is fitted on 90 percent of the samples, and the class of the remaining 10 percent is then predicted. This procedure is repeated 10 times, with each part playing the role of the test samples and the error of all 10 parts added together to compute the overall error. The error within the validation set reflects the number of samples wrongfully predicted to be in this set.

T7-(deoxythymidine)24-primer (Genset). In vitro transcription was performed with biotin-11-cytidine triphosphate and biotin-16-uridine triphosphate (Perkin–Elmer) and the MEGAScript T7 labeling kit (Ambion). Double-stranded cDNA and complementary RNA (cRNA) were purified and fragmented with the GeneChip Sample Cleanup module (Affymetrix). Biotinylated RNA was hybridized to the Affymetrix U133A GeneChip (45°C for 16 hours). Staining, washing, and scanning procedures were carried out as described in the GeneChip Expression Analysis technical manual (Affymetrix). All GeneChips were visually inspected for irregularities. The global method of scaling, or normaliza tion, was applied, and the mean (±SD) difference between the scaling, or normalization, factors of all GeneChips (293 samples; 285 from patients with AML, 5 from subjects with normal bone marrow,and 3 from subjects with CD34+ cell samples) was 0.70±0.26. All additional measures of quality — the percentage of genes present (50.6±3.8), the ratio of action 3′ to 5′ (1.24±0.19), and the ratio of *GAPDH* 3′ to 5′ (1.05±0.14) — indicated a high overall quality of the samples and assays. Detailed clinical, cytogenetic, and molecular cytogenetic information is available at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo, accession number GSE1159).

2.3.4 Data normalization, analysis, and visualization

All intensity values were scaled to an average value of 100 per GeneChip according to the method of global scaling, or normalization, provided in the Affymetrix Microarray Suite software, version 5.0 (MAS5.0). Since our methods reliably identify samples with an average intensity value of 30 or more but do not reliably discriminate values between 0 and 30, these values were set to 30. This procedure affected 31 percent of all intensity values, of which 64 percent were flagged as absent by the MAS5.0 software, 3 percent were flagged as marginal, and 33 percent were flagged as present according to the MAS5.0 software. For each probe set, the geometric mean of the hybridization intensities of all samples from the patients was calculated. The level of expression of each probe set in every sample was determined relative to this geometric mean and logarithmically transformed (on a base 2 scale) to ascribe equal weight to gene-expression levels with similar relative distances to the geometric mean. Deviation from the geometric mean reflects differential gene expression. The transformed expression data were subsequently imported into Omniviz software, version 3.6 (Omniviz), significance analysis of microarrays (SAM) software, version 1.21, and prediction analysis of microarrays (PAM) software, version 1.12.

2.3.5 Use of Pearson's Correlation and Visualization Tool

The Omniviz package was used to perform and visualize the results of unsupervised cluster analysis (an analysis that does not take into account external information such as the morphologic subtype or karyotype). Genes (probe sets) whose level of expression differed from the geometric mean (reflecting up- or down-regulation) in at least one patient were selected for further analysis. The clustering of molecularly recognizable specific groups of patients was investigated with each of the selected probe sets with the use of the Pearson's Correlation and Visualization tool of Omniviz (provided in Fig. B, C, D, E, F, G, and H in Supplementary Appendix 1, available with the full text of this article at www.nejm.org).

Characteristic	Value
Sex - no.(%)	
Male	137 (48)
Female	148 (52)
Age group - no. (%)	
<35yr	76 (27)
35 - 60 yr	177 (62)
≥60 yr	32 (11)
Age - yr	
Median	44
Range	15 - 78
White-cell count - x 10^{-3}/mm^3	
Median	28
Range	0.3 - 582
Banoe marrow blast count - %	
Median	66
Range	0 - 98
Platelet count - x 10^{-3}/mm^3	
Median	45
Range	3 - 931
French-American-British classification - no. (%)	
M0	6 (2)
M1	63 (22)
M2	66 (23)
M3	19 (7)
M4	53 (19)
M5	65 (23)
M6	3 (1)
Not determined	10 (4)
Cytogenetic abnormalities - no. (%)*	
t(15;17)	18 (6)
t(8;21)	22 (8)
inv(16)/t(16;16)	19 (7)
+8	26 (9)
+11	7 (2)
+21	2 (1)
-5	3 (1)

Characteristic	Value
Cytogenetic abnormalities - no. (%)* *continued*	
-5(q)	1 (<1)
-7	13 (5)
-7(q)	7 (2)
3(q)	6 (2)
t(6;9)	4 (1)
t(9;22)	2 (1)
t(11q23)	19 (7)
Complex karyotype (>3 chromosomal abnormalities)	11 (4)
Other abnormal karyotypes	60 (21)
Normal karyotype	119 (42)
Not determined	10 (4)
Molecular abnormalities - no. (%)	
Mutation	
FLT3 internal tandem duplication	78 (27)
FLT3 tyrosine kinase domain	33 (12)
N-*RAS*	26 (9)
K-*RAS*	9 (3)
CEBPa	17 (6)
Overexpression *EVI1*	23 (8)

* *All patients with a specific cytogenetic abnormality were included in the analysis, irrespective of the presence of additional abnormalities. A summary of the frequencies and percentages of the cytogenetic and molecular abnormalities for each of the assigned clusters can be found in Table Q of Supplementary Appendix 1 (available with the full text of this chapter at www.nejm.org). Some samples had more than one abnormality.*

Table 1. Clinical and molecular characteristics of the 285 patients with newly diagnosed AML.

2.3.6 The SAM Method

All supervised analyses were performed with the use of SAM software (19). A supervised analysis correlates gene expression with an external variable such as the karyotype or the duration of survival. SAM calculates a score for each gene on the basis of the change in expression relative to the SD of all 285 measurements. The criteria for identifying the top 40 genes for an assigned cluster were a minimal difference in gene expression between the assigned cluster and the other AML samples by a factor of 2 and a q value of less than 2 percent. The q value for each gene represents the probability that it is falsely called significantly deregulated.

2.3.7 The PAM Method

All supervised class-prediction analyses were performed by applying PAM software in R (version 1.7.1).20 The method of the nearest shrunken centroids identifies a subgroup of genes that best characterizes a predefined class. The prediction error was calculated by means of 10-fold cross validation (see the Glossary) within the training set (two thirds of the patients) followed by the use of a second validation set (one third of the patients). All genes identified by the SAM and PAM methods are listed in Supplementary Appendix 1 (Tables A1 to P1 and R).

2.3.8 Reverse-transcriptase polymerase chain reactions and sequence analyses

Reverse-transcriptase–polymerase-chain-reaction (RT-PCR) assays and sequence analyses for internal tandem duplication and tyrosine kinase domain mutations in *FLT3* and mutations in *NRAS*, *KRAS*, and *CEBPa*, as well as real-time PCR for *EVI1* were performed as described previously (8,9,21,22). AML samples of the clusters characterized by favorable cytogenetic characteristics (t(8;21), t(15;17), and inv(16)) were analyzed for the expression of fusion genes by real-time PCR (Supplementary Appendix 1).

2.3.9 Statistical analysis

Statistical analyses were performed with Stata Statistical Software, release 7.0. Actuarial probabilities of overall survival (with failure defined as death from any cause) and event-free survival (with failure defined as incomplete remission [set at day 1], relapse, or death during a first complete remission) were estimated according to the Kaplan–Meier method.

2.4 Results

2.4.1 Visual correlation of gene expression

All specimens of AML were classified into subgroups with the use of unsupervised ordering (i.e., without taking into account hematologic, cytogenetic, or other external information). Optimal clustering of these specimens was reached with the use of 2856 probe sets (a probe set consists of 10 to 20 oligonucleotides); 2856 sets represent 2008 annotated genes and 146 expressed-sequence tags, which are short sequences of unknown genes (Fig.1A and Table 2, and Fig. B, C, D, E, F, G, and H in Supplementary Appendix 1). Sixteen distinct groups of patients with AML were identified on the basis of strong similarities in gene-expression profiles. Figure 1A, a Pearson's correlation view, shows these clusters as red squares along the diagonal. A red rectangle indicates positive pairwise correlations (equality in gene expression between clusters) and a blue rectangle indicates negative pairwise correlations (inequality in gene expression between clusters) (Fig. 1A, and Fig. A in Supplementary Appendix 1). The final Omniviz Correlation View was adapted so that cytologic, cytogenetic, and molecular features were plotted directly adjacent to the original diagonal. This arrangement allowed the visualization of groups of patients with similar patterns of gene expression along with relevant clinical and genetic findings (Fig. 1B). Distinct clusters of t(8;21), inv(16), and t(15;17) were readily identified with 1692 probe sets (Table 2). Identification of clusters with mutations

in FLT3, monosomy 7, or overexpression of EVI1 required 2856 probe sets (Table 2, and Fig. B, C, D, E, F, G, and H in Supplementary Appendix 1). When more genes were used, the compact pattern of clustering vanished (Table 2). When included in the Omniviz Correlation View analyses (2856 probe sets), all five samples of bone marrow and three CD34+ samples from control subjects gathered within clusters 8 and 10, respectively. Genes characteristic of each of the 16 clusters were obtained by means of supervised analysis (distinctions on the basis of predefined classes), with the use of the SAM method. The expression profiles of the top 40 genes of each cluster are plotted in Figure 1B beside the correlation view. The SAM analyses identified 599 discriminating genes (Tables A1 to P1 in Supplementary Appendix 1); we were unable to identify a distinct gene profile for cluster 14.

2.4.2 Recurrent translocations

CBFβ-MYH11
All AML samples with inv(16), which causes the *CBFβ-MYH11* fusion gene, gathered within cluster 9 (Fig. 1B, and Table I in Supplementary Appendix 1). Four specimens within this cluster were not known to harbor an inv(16), but molecular analysis and Southern blotting revealed that their leukemic cells had the *CBFβ-MYH11* fusion gene (Table I and Fig. I in Supplementary Appendix 1). SAM analysis revealed that *MYH11* was the most discriminative gene for this cluster (Table I1 and Fig. J in Supplementary Appendix 1). Interestingly, a low level of expression of *CBFβ* was correlated with this cluster, perhaps because of the decreased expression or deletion of the *MYH11-CBFβ* alternate fusion gene or down-regulation of the normal *CBFβ* allele by the *CBFβ-MYH11* fusion protein.

PML-RARa
Cluster 12 contained all cases of acute promyelocytic leukemia (APL) with t(15;17) (Fig. 1B, and Table L in Supplementary Appendix 1), including one patient (Patient 322) who had previously received a diagnosis of APL with *PML-RARa* on the basis of RT-PCR alone. SAM analyses revealed that genes for hepatocyte growth factor (*HGF*), macrophage-stimulating 1 growth factor (*MST1*), and fibroblast growth factor 13 (*FGF13*) were specific for this cluster. In addition, cluster 12 could be separated into two subgroups: one with a high and the other with a low white-cell count (Fig. K in Supplementary Appendix 1). This subdivision corresponds to the presence of *FLT3* internal tandem duplication mutations (Fig. 1B).

AML1-ETO
All specimens from patients with the t(8;21) that generates the *AML1-ETO* fusion gene grouped within cluster 13 (Fig. 1B, and Table M in Supplementary Appendix 1). SAM identified *ETO* as the most discriminative gene for this cluster (Table M1 and Fig.L in Supplementary Appendix 1).

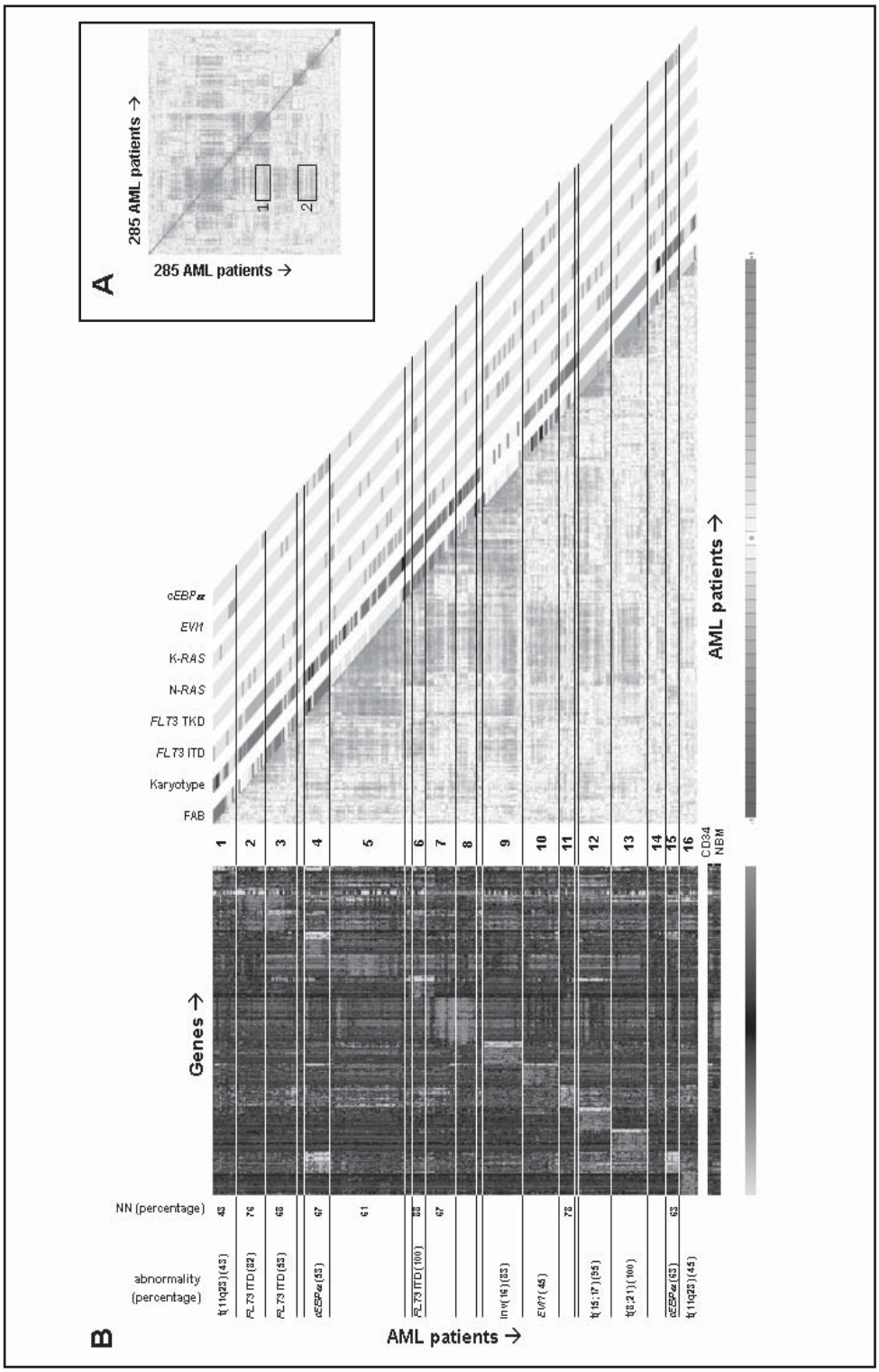
2.4.3 11q23 abnormalities

Cases with 11q23 abnormalities were scattered among the 285 samples, although two subgroups were apparent: cluster 1 and cluster 16 (Fig. 1B, and Tables A and P in Supplementary Appendix 1). Cluster 16, with 11 total cases, contained 4 cases

of t(9;11) and 1 case of t(11;19). SAM analyses identified a strong signature of up-regulated genes in most cases in this cluster (Fig. 1B, and Table P1 in Supplementary Appendix 1). Although 6 of 14 cases within cluster 1 also had 11q23 abnormalities, this subgroup was more heterogeneous than cluster 16 (Fig. 1B).

Figure 1 (facing page). Correlation View of Specimens from 285 Patients with AML Involving 2856 Probe Sets (Panel A) and an Adapted Correlation View (2856 Probe Sets) (Right-Hand Side of Panel B), and the Levels of Expression of the Top 40 Genes That Characterized Each of the 16 Individual Clusters (Left-Hand Side of Panel B). *In Panel A, the Correlation Visualization tool displays pairwise correlations between the samples. The colors of the cells relate to Pearson's correlation coefficient values, with deeper colors indicating higher positive (red) or negative (blue) correlations. One hundred percent negative correlation would indicate that genes with a high level of expression in one sample would always have a low level of expression in the other sample and vice versa. Box 1 indicates a positive correlation between clusters 5 and 9 and box 2 a negative correlation between clusters 5 and 12. The red diagonal line displays the intraindividual comparison of results for a patient with AML (i.e., 100 percent correlation). To reveal the patterns of correlation, we applied a matrix-ordering method to rearrange the samples. The ordering algorithm starts with the most highly correlated pair of samples and, through an iterative process, sorts all the samples into correlated blocks. Each sample is joined to a block in an ordered manner so that a correlation trend is formed within a block, with the most correlated samples at the center. The blocks are then positioned along the diagonal of the plot in a similar ordered manner. Panel B shows all 16 clusters identified on the basis of the Correlation View. The French-American-British (FAB) classification and karyotype based on cytogenetic analyses are depicted in the columns along the original diagonal of the Correlation View; FAB subtype M0 is indicated in black, subtype M1 in green, subtype M2 in purple, subtype M3 in orange, subtype M4 in yellow, subtype M5 in blue, andsubtype M6 in gray; normal karyotypes are indicated ingreen, inv(16) abnormalities in yellow, t(8;21) abnormalitiesin purple, t(15;17) abnormalities in orange, 11q23 abnormalities in blue, 7(q) abnormalities in red, +8 aberrations in pink, complex karyotypes (those involving more than three chromosomal abnormalities) in black, and other abnormalities in gray. FLT3 internal tandem duplication (ITD) mutations, FLT3 mutations in the tyrosine kinase domain (TKD), NRAS, KRAS, and CEBPa mutations, and the overexpression of EVI1 are depicted in the same set of columns: red indicates the presence of a given abnormality, and green its absence. The levels of expression of the top 40 genes identified by the significance analysis of microarrays of each of the 16 clusters as well as in normal bone marrow (NBM) and CD34+ cells are shown on the left side. The scale bar indicates an increase (red) or decrease (green) in the level of expression by a factor of at least 4 relative to the geometric mean of all samples. The percentages of the most common abnormalities (those present in more than 40 percent of specimens) and the percentages of specimens in each cluster with a normal karyotype are indicated.*

A full-color version of this figure is provided on the CD.


A
285 AML patients →
285 AML patients →
1
2
B
Genes →
AML patients →
AML patients →
cEBPα
EVI1
K-RAS
N-RAS
FLT3 TKD
FLT3 ITD
Karyotype
FAB
CD34
NBM
NN (percentage)
abnormality (percentage)
t(11q23) (43)
FLT3 ITD (62)
FLT3 ITD (55)
cEBPα (55)
FLT3 ITD (100)
inv(16) (85)
EVI1 (45)
t(15;17) (95)
t(8;21) (100)
cEBPα (62)
t(11q23) (45)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
43
76
68
67
61
88
67
73

Variable	Distribution						
No. of probe sets	147	293	569	984	1692	2856	5071
Factor increase or decrease in regulation†	>32	>22.6	>16	>11.3	>8	>5.6	>4
Chromosomal abnormalities							
t(8;21)	±	+	+	+	++	++	+
inv(16)	±	±	±	+	++	++	+
t(15;17)	±	+	++	++	++	++	+
11q23	±	±	±	±	+	+	±
-7(q)	±	±	±	±	±	+	±
Mutation							
FLT3 internal tandem duplication	±	±	±	±	±	±	±
FLT3 tyrosine kinase domain	-	-	-	-	-	-	-
N-*RAS*	-	-	-	-	-	-	-
K-*RAS*	-	-	-	-	-	-	-
CEBPa	-	±	±	+	+	+	+
Overexpression							
EVI1	-	-	-	-	±	+	±

** Two plus signs indicate that 100 percent of specimens were in a single cluster, a single plus sign that specimens were in no more than two recognizable clusters, a plus–minus sign that specimens were in more than two recognizable clusters, and a minus sign that no clustering occurred. Four patients with AML with abnormalities involving chromosome 5 were excluded.*

† The factor increase or decrease in the regulation of gene expression is relative to the geometric mean by which the differentially expressed probe sets were selected.

Table 2. Evaluation of the Omniviz Correlation View results on the basis of the clustering of AML specimens with similar molecular abnormalities.*

2.4.4 *CEBPa* mutations

Mutations in *CEBPa* occur in approximately 7 percent of patients with AML, most with a normal karyotype, and predict a favorable outcome (9,10). Two clusters (4 and 15) had a high frequency of *CEBPa* mutations (Fig. 1B). The sets of up-regulated or down-regulated genes in cluster 4 discriminated the specimens it contained from those in cluster 15 (Table D1 in Supplementary Appendix 1). The upregulated genes included the T-cell genes *CD7* and the T-cell receptor delta locus, which may be ex- pressed by immature AML cells. (23,24). All but one of the top 40 genes of cluster 15 were down-regulated (Table O1 in Supplementary Appendix 1). These genes were also down-regulated in cluster 4 (Fig. 1B). The genes encoding alpha1-catenin (*CTNNA1*), tubulin beta-5 (*TUBB5*), and Nedd4 family interacting protein

1 (*NDFIP1*) were the only down-regulated genes among the top 40 in both cluster 4 and cluster 15.

2.4.5 Overexpression of *EVI1*

High levels of expression of *EVI1*, which occur in approximately 10 percent of cases of AML, predict a poor outcome (8). In cluster 10, 10 of 22 specimens (Table J in Supplementary Appendix 1) showed increased expression of *EVI1*, and 6 of these 10 specimens had chromosome 7 abnormalities. In cluster 8, 4 of 13 specimens also had chromosome 7 aberrations (Table H in Supplementary Appendix 1), but since its molecular signature differed from that of cluster 10 (Fig. 1B), the high level of expression of *EVI1* or *EVI1*-related proteins may have determined the molecular profile of cluster 10. In the heterogeneous cluster 1, 5 of 14 specimens also had increased *EVI1* expression. These specimens may have appeared outside cluster 10 because their molecular signatures were most likely the result of the overexpression of *EVI1* and an 11q23 abnormality.

2.4.6 *FLT3* and *RAS* mutations

Samples from most patients in clusters 2, 3, and 6 harbored a *FLT3* internal tandem duplication (Fig.1B). Almost all these patients had a normal karyotype. The presence of *FLT3* internal tandem duplication seemed to divide clusters 3, 5, and 12 into two groups. Other individual specimens with a *FLT3* internal tandem duplication were dispersed over the entire series; mutations in the tyrosine kinase domain of *FLT3* were not clustered. Likewise, mutations in codon 12, 13, or 61 of the small GTPase RAS (*N-RAS* and *K-RAS*) had no apparent signatures and did not aggregate in the Correlation View (Fig. 1B).

2.4.7 Other clusters

Specimens from patients with AML with a normal karyotype clustered into several subgroups within the assigned clusters (Fig. 1B). Most patients in cluster 11 had normal karyotypes and no consistent dditional abnormality. Cluster 5 contained mainly specimens from patients with AML of subtype M4 or M5, according to the French–American–British (FAB) classification (Fig. 1B). Clusters 7, 8, 11, and 14 were not associated with a FAB subtype but had distinct gene-expression profiles.

2.4.8 Class prediction of distinct clusters

We used the PAM method to validate the cluster specific genes identified by the SAM method and to determine the minimal number of genes that can be used to predict karyotypic or other genetic abnormalities with biologic significance in AML (Table 3). The 285 specimens were randomly divided into a training set (189 specimens) and a validation set (96 specimens). All patients in the validation set who had favorable cytogenetic findings were identified with 100 percent accuracy with the use of only a few genes (Table 3). As expected from the SAM analyses, *ETO* for t(8;21), *MYH11* for inv(16), and *HGF* for t(15;17) were among the best predictors of the cytogenetic abnormalities (Table R in Supplementary Appendix 1). Cluster 10 (which involved *EVI1* overexpression) was predicted with a high degree of accuracy, although with a higher 10-fold cross-validation error than that

Abnormality	Training Set (n=189)	Validation Set (n=96)	No. of probe sets used	No. of genes represented
	no.of errors			
t(8;21), leading to *AML1-ETO* (cluster 13)	0	0	3	2
t(15;17), leading to *PML-RARa* (cluster 12)	1	0	3	2
inv(16), leading to *CBFβ-MYH11* (cluster 9)	0	0	1	1
11q23 (cluster 16)	3	3	31	25
EVI1 (cluster 10)	16	0	28	25
CEBPa (cluster 4)	8	2	13	8
CEBPa (cluster 15)	17	6†	36	32
CEBPa (cluster 4 and 15)	5	2	9	5
FLT3 internal tandem duplication	27	21	56	41

** Prediction analysis of microarrays was performed to define the minimal numbers of genes that could predict whether a specimen from a particular patient belonged in one of the clusters (first column). The group of patients was randomly segregated into a training set (second column) and a validation set (third column). The 10-fold method of cross-validation, applied on the training set, works as follows: the model is fitted on 90 percent of the samples, and the class of the remaining 10 percent is then predicted. This procedure is repeated 10 times, with each part playing the role of the test samples and the error of all 10 parts added together to compute the overall error (second column). The minimal numbers of probe sets or genes (fourth and fifth columns, respectively) that were identified in the training were tested on the validation set (third column). The error within the validation set (third column) reflects the number of samples wrongfully predicted in this set. The identities of the probe sets and genes are provided in Table R of Supplementary Appendix 1.*

† After randomization none of the patients with CEBPa abnormalities in cluster 15 were included in the validation set.

Table 3. Results of class prediction analysis with the use of prediction analysis of microarrays.

in the groups with favorable cytogenetic findings. In cluster 16 (involving 11q23 abnormalities), samples from 3 of 96 patients were wrongfully identified in the validation set. Since cluster 15 (involving *CEBPa* mutations) contained few samples, we combined both *CEBPa*-containing clusters. These combined clusters predicted the presence of *CEBPa* mutations within the validation set with 98 percent accuracy.

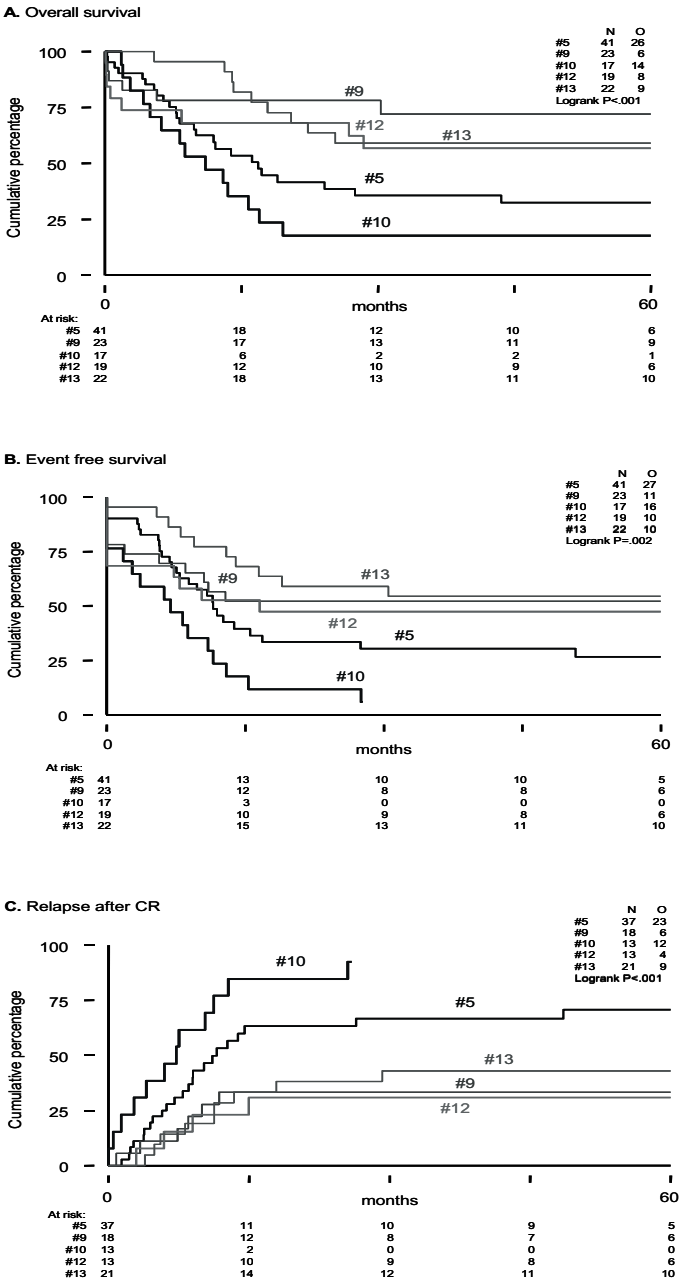


Figure 2. Kaplan–Meier Estimates of Overall Survival (Panel A), Event-free Survival (Panel B), and Relapse Rates after Complete Remission (Panel C) among Patients with AML with Specimens in Clusters 5, 9, 10, 12, and 13. *Cluster 5 was characterized by a French–American–British classification of M4 or M5, cluster 9 by inv(16) abnormalities, cluster 10 by a high level of expression of EVI1, cluster 12 by t(15;17) abnormalities, and cluster 13 by t(8;21) abnormalities. P values were calculated with the use of the log-rank test.*

We were unable to identify a signature that reliably identified *FLT3* internal tandem duplications.

2.4.9 Survival analyses

Overall survival, event-free survival, and relapse rates were determined among patients whose specimens were within clusters containing more than 20 specimens in the Correlation View (clusters 5, 9, 10, 12, and 13) (Fig. 2). The mean (±SE) actuarial probabilities of overall survival and event-free survival at 60 months were 59±10 percent and 55±11 percent, respectively, among patients with samples in cluster 13; 57±12 percent and 47±11 percent, respectively, among those with samples in cluster 12; and 72±10 percent and 52±10 percent, respectively, among those with samples in cluster 9. Patients with samples in cluster 5 had an intermediate rate of overall survival (32±8 percent) and event-free survival (27±8 percent), whereas survival among patients with samples in cluster 10 was poorer (the overall survival rate was 18±9 percent, and the event free survival rate was 6±6 percent), mainly as a result of an increased incidence of relapse (Fig. 2C).

2.5 *Discussion*

In this study of 285 patients with AML that was characterized by cytogenetic analyses and extensive molecular analyses, we used gene-expression profiling to comprehensively classify the disorder. This method identified 16 groups on the basis of unsupervised analyses involving Pearson's correlation coefficient. Our results provide evidence that each of the assigned clusters represents true subgroups of AML with specific molecular signatures.

We were able to cluster all cases of AML with t(8;21), inv(16), or t(15;17), including those that had not been identified by cytogenetic examination, into three clusters with unique gene-expression profiles. Correlations between gene-expression profiles and prognostically favorable cytogenetic aberrations have been reported by others, (12,13) but we found that these cases can be recognized with a high degree of accuracy within a representative cohort of patients with AML.

The SAM and PAM methods were highly concordant for the genes identified within the assigned clusters, indicating that these clusters contained discriminative genes. For instance, clusters 4 and 15, with overlapping signatures, both included specimens with normal karyotypes and mutations in *CEBPa*. Multiple genes appeared to be down-regulated in both clusters but were unaffected in any other subgroup of AML.

The discriminative genes identified by SAM and PAM may reveal functional pathways that are critical for the development of AML. These methods of statistical treatment of the data identified several genes that are implicated in specific subtypes of AML, such as the interleukin-5 receptor a (*IL5Ra*) gene in AML with t(8;21) abnormalities (25) and FLT3-STAT-5 targets — the gene for interleukin-2 receptora (*IL2Ra*) (26) and the pim1 kinase gene (*PIM1*) (27)— in AML with *FLT3* internal tandem duplication mutations.

Five clusters (5, 9, 10, 12, and 13) with 20 or more specimens were evaluated in relation to outcome of disease. As expected, clusters 9 (involving *CBFβ-MYH11*), 12 (involving *PML-RARa*), and 13 (involving *AML1-ETO*) contained specimens with a relatively favorable prognosis.

Specimens in cluster 10 had a distinctly poor outcome. A randomly selected subgroup of patients with specimens in this cluster could be identified with a high degree of accuracy with the use of a minimal number of genes. The high frequency of poor prognostic markers in this cluster (-7(q), -5(q), t(9;22), or high levels of expression of *EVI1*) is in accord with the poor outcome of patients in this cluster. Since this cluster is heterogeneous with regard to both known poor-risk markers and the presence or absence of these markers, the molecular signature of this cluster may signify a biochemical pathway that causes a poor outcome. The fact that normal CD34+ cells segregate into this cluster suggests that the molecular signature of treatment resistance resembles that of normal hematopoietic stem cells.

The 44 patients with specimens in cluster 5 had an intermediate duration of survival. Since these specimens were of the FAB M4 or M5 subtype, it is possible that genes related to monocytes or macrophages were important in the clustering of these cases.

In three clusters more than 75 percent of specimens had a normal karyotype (clusters 2, 6, and 11). Most of the patients with specimens in clusters 2 and 6 had *FLT3* internal tandem duplication mutations, whereas patients with specimens in cluster 11, which had a discriminative molecular signature, did not have any consistent molecular abnormality.

Clusters 1 and 16 harbored 11q23 abnormalities, representing defects involving the mixed-lineage leukemia (*MLL*) gene. The different gene-expression profiles of these two clusters are most likely due to additional distinctive genetic defects. In cluster 1, this additional abnormality may be a high level of expression of the oncogene *EVI1*, which was not apparent in cluster 16. Similarly, distinctive additional genetic defects may explain the separation of clusters 4 and 15, both of which contained specimens with *CEBPa* mutations, clusters 1 and 10, both of which had high levels of *EVI1* expression, and clusters 8 and 10, both of which had a high frequency of monosomy 7.

Internal tandem duplications in *FLT3* adversely affect the clinical outcome (6,7). The molecular signature associated with this abnormality is not distinctive; however, the clustering of specimens with these abnormalities within assigned clusters (e.g., cluster 12) suggests that these internal tandem duplications result in different biologic entities within the scope of AML.

Our study demonstrates that cases of AML with known cytogenetic abnormalities and new clusters of AML with characteristic gene-expression signatures can be identified with the use of a single assay. The applicability and performance of genome-wide analysis will advance with the availability of novel whole-genome arrays, improved sequence annotation, and the development of sophisticated protocols and software, allowing the analysis of subtle differences in gene expression and predictions of pathogenic pathways.

We are indebted to Gert J. Ossenkoppele, M.D. (Free University Medical Center, Amsterdam), Edo Vellenga, M.D. (University Hospital, Groningen, the Netherlands), Leo F. Verdonck, M.D. (University Hospital, Utrecht, the Netherlands), Gregor Verhoef, M.D. (Hospital Gasthuisberg, Leuven, Belgium), and Matthias Theobald, M.D. (Johannes Gutenberg University Hospital, Mainz, Germany), for providing AML samples; to our colleagues from the bone marrow transplantation group and molecular diagnostics laboratory for

storing the samples and performing the molecular analyses, respectively; to Guang Chen (Omniviz, Maynard, Mass.); to Elisabeth M.E. Smit (Erasmus Medical Center, Rotterdam, the Netherlands) for cytogenetic analyses; to Wim L.J. van Putten, Ph.D. (Erasmus Medical Center, Rotterdam, the Netherlands), for statistical analyses; to Ivo P. Touw, Ph.D. (Erasmus Medical Center, Rotterdam, the Netherlands), for helpful discussions; and to Eveline Mank (Leiden Genome Technology Center, Leiden, the Netherlands) for initial technical assistance.

References

1. Lowenberg B, Downing JR, Burnett A. Acute myeloid leukemia. N Engl J Med 1999, 341(14):1051-1062.
2. Slovak ML, Kopecky KJ, Cassileth PA, et al: Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group Study. Blood 2000, 96(13):4075-4083.
3. Byrd JC, Mrozek K, Dodge RK, et al. Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). Blood 2002, 100(13):4325-4336.
4. Grimwade D, Walker H, Oliver F, et al. The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. Blood 1998, 92(7):2322-2333.
5. Grimwade D, Walker H, Harrison G, et al. The predictive value of hierarchical cytogenetic classification in older adults with acute myeloid leukemia (AML): analysis of 1065 patients entered into the United Kingdom Medical Research Council AML11 trial. Blood 2001, 98(5):1312-1320.
6. Gilliland DG, Griffin JD: The roles of FLT3 in hematopoiesis and leukemia. Blood 2002, 100(5):1532-1542.
7. Levis M, Small D: FLT3: ITDoes matter in leukemia. Leukemia 2003, 17(9):1738-1752.
8. Barjesteh van Waalwijk van Doorn-Khosrovani S, Erpelinck C, van Putten WL, et al. High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. Blood 2003, 101(3):837-845.
9. 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 2002, 100(8):2717-2723.
10. van Waalwijk van Doorn-Khosrovani SB, 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.
11. Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002, 30(1):41-47.
12. Debernardi S, Lillington DM, Chaplin T, et al. Genome-wide analysis of acute myeloid leukemia with normal karyotype reveals a unique pattern of homeobox gene expression distinct from those with translocation-mediated fusion events. Genes Chromosomes Cancer 2003, 37(2):149-158.
13. Schoch C, Kohlmann A, Schnittger S, et al. Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. Proc Natl Acad Sci U S A 2002, 99(15):10008-10013.
14. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286(5439):531-537.
15. 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 1997, 15(12):3496-3506.
16. 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 2003, 349(8):743-752.
17. 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 elderly AML. Blood 2004, Apr 15;103(8):2908-13.
18. Chomczynski P, Sacchi N: Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 1987, 162(1):156-159.
19. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001, 98(9):5116-5121.
20. Tibshirani R, Hastie T, Narasimhan B, Chu G: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002, 99(10):6567-6572.
21. Valk PJM, Bowen DT, Frew ME, Goodeve AC, Löwenberg B, Reilly JT: Second hit mutations in the RTK/RAS signalling pathway in acute myeloid leukaemia and inv(16). Haematologica 2004, 89(01):106.
22. Care RS, Valk PJ, Goodeve AC, et al. Incidence and prognosis of c-KIT and FLT3 mutations in core binding factor (CBF) acute myeloid leukaemias. Br J Haematol 2003, 121(5):775-777.
23. Lo Coco F, De Rossi G, Pasqualetti D, et al. CD7 positive acute myeloid leukaemia: a subtype associated with cell immaturity. Br J Haematol 1989, 73(4):480-485.
24. Boeckx N, Willemse MJ, Szczepanski T, van der Velden VH, Langerak AW, Vandekerckhove P, van Dongen JJ: Fusion gene transcripts and Ig/TCR gene rearrangements are complementary but infrequent targets for PCR-based detection of minimal residual disease in acute myeloid leukemia. Leukemia 2002, 16(3):368-375.
25. Touw I, Donath J, Pouwels K, et al. Acute myeloid leukemias with chromosomal abnormalities involving the 21q22 region identified by their in vitro responsiveness to interleukin-5. Leukemia 1991, 5(8):687-692.
26. Kim HP, Kelly J, Leonard WJ: The basis for IL-2-induced IL-2 receptor alpha chain gene regulation: importance of two widely separated IL-2 response elements. Immunity 2001, 15(1):159-172.
27. Lilly M, Le T, Holland P, Hendrickson SL: Sustained expression of the pim-1 kinase is specifically induced in myeloid cells by cytokines whose receptors are structurally related. Oncogene 1992, 7(4):727-732.

Chapter 3

The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies

Roel G.W. Verhaak[1], Frank J.T. Staal[2], Peter J.M. Valk[1], Bob Löwenberg[1], Marcel J.T. Reinders[3] and Dick de Ridder[2,3]

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

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

[3]Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, the Netherlands

BMC Bioinformatics. 2006 Mar 2;7:105.

3.1 Abstract

3.1.1 Background

Intensity values measured by Affymetrix microarrays have to be both normalized, to be able to compare different microarrays by removing non-biological variation, and summarized, generating the final probe set expression values. Various pre-processing techniques, such as dChip, GCRMA, RMA and MAS have been developed for this purpose. This study assesses the effect of applying different pre-processing methods on the results of analyses of large Affymetrix datasets. By focusing on practical applications of microarray-based research, this study provides insight into the relevance of pre-processing procedures to biology-oriented researchers.

3.1.2 Results

Using two publicly available datasets, i.e., gene-expression data of 285 patients with Acute Myeloid Leukemia (AML, Affymetrix HG-U133A GeneChip) and 42 samples of tumor tissue of the embryonal central nervous system (CNS, Affymetrix HuGeneFL GeneChip), we tested the effect of the four pre-processing strategies mentioned above, on (1) expression level measurements, (2) detection of differential expression, (3) cluster analysis and (4) classification of samples. In most cases, the effect of pre-processing is relatively small compared to other choices made in an analysis for the AML dataset, but has a more profound effect on the outcome of the CNS dataset. Analyses on individual probe sets, such as testing for differential expression, are affected most; supervised, multivariate analyses such as classification are far less sensitive to pre-processing.

3.1.3 Conclusion

Using two experimental datasets, we show that the choice of pre-processing method is of relatively minor influence on the final analysis outcome of large microarray studies whereas it can have important effects on the results of a smaller study. The data source (platform, tissue homogeneity, RNA quality) is potentially of bigger importance than the choice of pre-processing method.

3.2 Introduction

The analysis of gene expression data generated by microarrays, such as the high-density oligonucleotide microarays produced by Affymetrix (Santa Clara, CA), is an often laborious process in which a basic understanding of molecular biology, computer science and statistics is required. In a typical microarray experiment, RNA obtained under various conditions (patients, treatments, disease states etc.) is hybridised to microarrays. By tagging the RNA with a fluorescent marker, intensity values can be obtained that correspond to the amount of labeled RNA bound to the array. On the widely used Affymetrix platform, gene expression is measured using probe sets consisting of 11 to 20 perfect match (PM) probes of 25 nucleotides, which are complementary to a target sequence, and a similar number of mismatch (MM) probes in which the 13th nucleotide has been changed. The MM probe measurements are thought to comprise most of the background cross-hybridization and stray signal affecting the PM probes.

Normalization of probe intensity values is performed to remove any non-biological variation. The individual probe measurements are then summarized as probe set expression levels, as estimates of the amount of specific mRNA present in the biological sample. Normalization and probe set summarization are statistical procedures for which several methods have been developed. MicroArray Suite (MAS 5.0), a software package provided by Affymetrix, normalizes intensities using a global scaling procedure and measures expression using a one-step Tukey biweight algorithm, which is defined as the anti-log of a robust average of differences between log(PM) and log(MM) (1). The same algorithms are implemented in the software package currently provided by Affymetrix, GCOS. One of the first alternatives to this approach was provided by Li and Wong with the dChip-method, which scales the intensity data towards the median intensity in a group of arrays and then uses model-based index estimates, giving variable weight to PM-MM probe pairs of a probe set based on variance between arrays, to measure expression (2). Irizarry et al. introduced RMA (robust multi-array average), later followed by GCRMA (GC robust multi-array average). RMA, often preceded by quantile normalization (3, 4), applies a median polish procedure to PM intensities only in summarization. GCRMA is based on a similar model as RMA but takes into account the effect of stronger bonding of G/C pairs (5, 6). An overview of these methods is shown in Table 1. Other normalization methods, such as the variance stabilizing normalization (VSN, (7)) and summarization methods, such as PLIER (8), have been developed, but are less frequently applied.

Various studies have been published which assess the differences in outcome of these different data pre-processing methods (9-14). To validate and test pre-processing methods, two publicly available datasets are commonly used. The Latin square dataset provided by Affymetrix (http://www.affymetrix.com/) contains spiked-in cRNA's at several concentrations facilitating the assessment of the relation between mRNA concentration and expression value. The GeneLogic dilution series (obtainable on request, http://www.genelogic.com) gives an estimate of the relation between actual and measured differential expression. Based on these datasets, an online benchmark tool has been developed to encourage authors to test their method (http://affycomp.biostat.jhsph.edu) (15). This tool assesses quality of pre-processing using several parameters in five different groups: (1) variability of expression across replicate arrays, (2) response of expression measure to changes in abundance of RNA, (3) sensitivity of fold-change measures to the amount of actual RNA sample, (4) accuracy of fold-change as a measure of relative expression and (5) usefulness of raw fold-change score for the detection of differential expression. Pronounced differences between different procedures have been shown to occur (4, 9, 11, 13).

The studies on the Latin square and dilution data were performed using data generated specifically for this purpose, allowing comparisons of specific analyses, showing accurately which methods perform best. In effect, statistical properties of the various estimators are tested. Several authors noted that the use of two special-purpose datasets for calibration of statistical procedures creates a risk on overfitting of the available data and therefore focused on using experimental data to compare methods with respect to the sets of differentially expressed genes found (9, 11, 14). This sometimes lead to contradictory results, where for instance a

	Normalization method	Summarization method
MAS	Global scaling – individual array normalization; moderate influence on expression levels, no effect on outliers. Non-parametric methods are potentially more reliable	Tukey biweight (robust average) – subtract MM from PM and adjust for negative values
dChip	Average median scaling - individual array normalization; moderate influence on expression levels, no effect on outliers. Non-parametric methods are potentially more reliable	Model-based index estimate – subtract MM from PM, but take indivual probe variability, assessed over all available arrays, into account
RMA	Quantile normalization – multiple array normalization; considerable influence on expression levels, with removal of outliers. Parametric methods are potentially more reliable	Median polish – only use MM for background adjustment; fit parameters of linear model robustly using median polish, taking into account all available arrays
GCRMA	Quantile normalization – multiple array normalization; considerable influence on expression levels, with removal of outliers. Parametric methods are potentially more reliable	Median polish - only use MM for background adjustment; fit parameters of linear model robustly using median polish, taking into account all available arrays; fit extra GC-content parameter

Table 1. Overview of several pre-processing methods.

study using the Latin square dataset showed the MAS5.0 method to outperform the dChip method on detecting differentially expressed genes (13), while a study on experimental data showed the opposite (9). Therefore, more work is needed to reliably establish how important the effect of choice of pre-processing method is in every-day practice, especially when analyses such as clustering and classification are applied.

In this paper, we focus on one practical application of microarrays: patient-cohort studies (16-20). In such studies, researchers typically select sets of genes that are differentially expressed between certain known conditions, a supervised analysis. Moreover, unsupervised techniques (not imposing any prior knowledge on the data) such as clustering are applied to detect biological relations between samples or genes by grouping them according to their expression profiles. Often the goal is to obtain a predictor (classifier) for, for instance, prognostically relevant categories, using supervised analysis.

Given that different pre-processing procedures will influence the outcome of these analyses, several questions can be asked, such as: How well is expression measured using a number of different pre-processing methods? What is their effect on the detection of differentially expressed genes, clusters found and classification

results? By focusing on practical applications of microarray studies, we hope to give insight into the relevance of different pre-processing procedures to biology-oriented researchers.

3.3 Materials and methods

3.3.1 Datasets

The two datasets used have been described before (17, 20). The first dataset consists of microarray measurements taken on samples of 285 patients with acute myeloid leukemia (AML), of whom blasts and mono-nuclear cells were isolated from peripheral blood or bone marrow aspirates. The samples were hybridized on Affymetrix HG-U133A GeneChip microarrays. This dataset will be referred to as the AML dataset; it is available on the Gene Expression Omnibus website (http://ncbi.nlm.nih.gov/geo, accession number GSE1159). The second dataset contains gene-expression data of 42 homogenized tumor tissues of the embryonal central nervous system (CNS), hybridized on Affymetrix HuGeneFL arrays. The dataset, referred to as the CNS dataset, is available at http://www.genome.wi.mit.edu/MPR/CNS/.

3.3.2 Normalization and expression measurement

Both datasets were pre-processed with MAS, RMA, GCRMA and dChip, resulting in eight different datasets. MAS expression data combined with global scaling was obtained from the MAS 5.0 software, provided by Affymetrix (Affymetrix Inc., Santa Clara, CA). dChip pre-processing together with scaling of the data towards the median average expression value per chip was applied using software available from the authors (http:///www.dchip.org/). RMA and GCRMA pre-processing was performed together with quantile normalization using the Bioconductor v2.0 library available in the R software environment (http://www.bioconductor.org/).

3.3.3 Real-time quantitative PCR (RQ-PCR)

For the AML dataset only, a number of measured probe set expression levels were compared to available RQ-PCR measurements of the corresponding genes on subsets of the original dataset (with n varying between 208 and 277, as indicated in Supp. Table 2). Probe sets were selected for RQ-PCR measurement based on biological relevance to the study of leukemia; samples were selected based on availability of material. Eligible patients had a diagnosis of primary AML, confirmed by cytological examination of blood and bone marrow. After informed consent, bone marrow aspirates or peripheral blood samples were taken at diagnosis. Blasts and mononuclear cells were purified by Ficoll-Hypaque (Nygaard, Oslo, Norway) centrifugation and cryopreserved. The AML samples contained 80-100 percent blast cells after thawing, regardless of the blast count at diagnosis.
After thawing, cells were washed once with Hanks balanced salt solution. High quality total RNA was extracted by lysis with guanidinium isothiocyanate followed by cesium chloride gradient purification. RNA concentration, quality and purity were examined using the RNA 6000 Nano assay on the Agilent 2100 Bioanalyzer (Agilent, Amstelveen, The Netherlands). None of the samples showed RNA

degradation (28S/18S rRNA ratio ≥ 2) or contamination by DNA.
cDNA was synthesized from 1µg of RNA using random hexamer priming, essentially as described (21). cDNA prepared from 50ng of RNA was used for all RQ-PCR amplifications.
Real-time quantitative PCR amplification was performed with the ABI PRISM 7700 sequence Detector (Applied Biosystems, Nieuwerkerk aan den IJssel, Netherlands), using 50 µL mix containing 20 µM deoxyribonucleoside triphosphates (dNTPs; Amersham Pharmacia Biotech, Roosendaal, Netherlands); 15 pmol forward and reverse primer (Life Technologies); 3 mM MgCl2 (5 mM for the reference gene porphobilinogen deaminase [PBGD]); 10 pmol probe (Eurogentec, Maastricht, Netherlands); 5 µL 10 x buffer A and 1.25 U AmpliTaq Gold (Applied Biosystems). The primers and probe combinations for detection of *EVI1* [EMBL:BX647613] (22), *CEBPa* [RefSeq:NM_004364.2] (23), *TRKA* [EMBL:M23102] (24) and *PBGD* [EMBL: AB162702] (24) have been described. Primer and probe combinations used to determine the expression of *MEIS1* [EMBL:AB040810] , *HOXA7* [EMBL:AJ005814], *PRDM1* [EMBL:AL358952], and *PRDM2* [EMBL:U23736] are listed in Supp. Table 1. Expression of *HOXA9* [EMBL:BC006537], *GMCSF* [EMBL:X03021], *P8* [EMBL: AF135266] was measured with 1x SYBR Green I dye (Applied Biosystems). The primers used in the SYBR Green reactions are listed in Supp. Table 1. The thermal cycling conditions included 10 minutes at 95°C followed by 45 cycles of denaturation for 30 seconds at 95°C and annealing/extension at 60°C for 60 seconds.
To quantify the relative expression levels of the various genes in AML the Ct values were normalized for the endogenous reference PBGD ($\Delta Ct = Ct_{target} - Ct_{PBGD}$) and compared with a calibrator NBM cells from healthy volunteers, using the $\Delta\Delta Ct$ method ($\Delta\Delta Ct = \Delta Ct_{AML\ sample} - \Delta Ct_{Calibrator}$). We used the $\Delta\Delta Ct$ value to calculate relative expression ($2^{-\Delta\Delta Ct}$).
A minimum threshold of 1 was applied, as well as log(2) transformation (25). Pearson correlation coefficients were calculated between the RQ-PCR data and the corresponding microarray-data pre-processed by the different procedures. Pearson correlation coefficients between data from the different procedures were also calculated, for each probe set present on the microarray.

3.3.4 Data transformation

For each probe set, the geometric mean m of all expression values e over the different samples was calculated. The level of expression for a particular sample was subsequently determined as $\log_2(e)-\log_2(m)$. This transformation was applied to all datasets and only transformed data was used for detection of differential expression, cluster analysis and classification.

3.3.5 Differential expression

Tests for differential expression were performed on several biologically relevant groups, by comparing samples from a group to the remainder of the samples. Four groups were tested in the AML dataset: (1) samples with a recurrent mutation in the *FLT3* gene (n = 78), (2) samples with inversion of chromosome 16, (inv(16), n = 23), (3) samples with translocation of chromosomes 15 and 17 (t(15;17), n = 19) and (4) samples with translocation of chromosomes 8 and 21 (t(8;21), n = 22). In the CNS dataset, four groups were tested as well: (1) samples with primitive neuro-

ectodermal tumors (PNET, n = 8), (2) samples with medullablastomas (MED, n = 10), (3) samples with rhabdoid tumors (RHAB, n = 10) and (4) samples with malignant gliomas (GLIO, n = 10).
Student's t-test and Wilcoxon's rank sum test were applied to each probe set (26). The resulting p-values were adjusted for multiple testing by Šidák step-down adjustment to control the Family-Wise Error Rate or FWER (27). The Significance Analysis of Microarrays (SAM) permutation algorithm (Excel-version 1.21, available from http://www-stat.stanford.edu/~tibs/SAM/), controlling the False Discovery Rate (FDR), was also applied (28). SAM provides an estimate of the FDR known as a q-value. Each test was applied and lists of probe sets, considered significantly differentially expressed at an FWER or FDR of 5%, were retrieved. For all possible combinations (i.e. MAS-dChip, MAS-RMA, MAS-GCRMA, dChip-RMA, dChip-GCRMA and RMA-GCRMA) probe sets marked as significantly differentially expressed by both methods were counted. To be able to compare different combinations, an overlap ratio $R(A,B)$ was calculated between the number of probe sets detected as differentially expressed in both datasets A and B and the total number of unique probe sets detected in the two datasets:

$$R(A,B) = \frac{2p}{a+b}$$

where p is the number of probe sets significant in both datasets, a is the number of significant probe sets found in dataset A and b is the number of significant probe sets found in dataset B.

3.3.6 Cluster analysis

Subsets of n probe sets (for the AML dataset, n = 3000; for CNS, n = 1000) were created by ranking probe sets by their standard deviation over all samples, and selecting the top n. Samples in all datasets (4 pre-processing methods) were clustered using k-means clustering and hierarchical clustering on both correlation distance matrices, in which the distance between two samples x and y is defined as $1-\rho_{xy}$; and Euclidean distance matrices, as used in (17). Hierarchical clustering was performed using single, average and complete linkage. To be able to compare all methods and datasets, the number of clusters was fixed to the expected number of groups based on biological characteristics of the patient population, which was 12 for the AML dataset and 5 for the CNS dataset. To investigate the influence of this setting, the AML dataset was also clustered into 2 and 20 clusters and the CNS dataset was clustered into 2 and 10 clusters, respectively. During each run of the k-means algorithm it was randomly restarted 1000 times, retaining the solution yielding minimum cluster within-scatter, in an attempt to avoid local minima.
Clustering results were compared using the Jaccard index. The Jaccard-index $J(C_1,C_2)$ compares two clusterings C_1 and C_2 based on the number of similar sample pairs available in the clusters and results in a value between 0 (no similar pairs) and 1 (all pairs are equal). It is estimated as

$$J(C_1,C_2) = \frac{n_{12}}{n_{12}+n_1+n_2}$$

where n_{12} denotes the number of pairs of samples in the same cluster in C_1 and assigned to the same cluster in C_2, n_1 denotes the number of pairs in the same cluster in C_1, but in different clusters in C_2 and n_2 denotes the number of pairs in the same cluster in C_2, but in a different cluster in C_1.

The raw Jaccard index should be interpreted in the light of how stable C_1 and C_2 actually are. If a clustering C, obtained using a certain pre-processing method, changes when one or a few samples are removed, it is to be expected that using a different pre-processing method will also have an impact. To estimate stability, for each pre-processing method 100 pairs of random subsets each containing 90% of the samples were clustered. Each individual subset was transformed as described and n = 3000 (or n = 1000 for the CNS dataset) probe sets were selected (these sets were 97.1% identical on average). In each pair, both subsets were then clustered, and the Jaccard index between these two clusterings was calculated using the samples present in both subsets. This resulted in 100 Jaccard indices, giving an impression of the variability due to transformation and subset selection. Finally, normal distributions were fitted to the 100 Jaccard indices found.

For a Jaccard index resulting from a comparison between two pre-processing methods M_1 and M_2, the cumulative distribution function (CDF) of the normal distribution for both pre-processing methods is used to arrive at two stability-normalized Jaccard indices J_1^{SN} and J_2^{SN}. Figure 3A illustrates this. A value of 0.5 for J_i^{SN} (in Figure 3A obtained at a Jaccard index of 0.48 for MAS or 0.62 for RMA) indicates that differences between pre-processing methods fall well within the range of clustering variability for pre-processing method M_i; values higher than that indicate that clustering differences due to pre-processing are in fact smaller than the average differences between clusterings on subsampled datasets. Although the notion of stability has been used before in clustering (e.g. (29)), we believe this normalized index to be novel.

Note that for the *k*-means algorithm, stability-normalised Jaccard indices are displayed in Figures 3B and 3C as mean and standard deviation over 10 runs of the algorithm, each run the result of 1000 restarts (see above).

3.3.7 Classification

Three two-class problems were defined on the AML dataset: (1) samples with inversion of chromosome 16 (inv(16)) vs. all others, (2) samples with a mutation in the *FLT3*-gene vs. all others and (3) samples that showed continuous complete remission (CCR) vs. samples that did not. These problems were selected in increasing order of expected difficulty. In the case of the CNS dataset, four two-class problems (PNET vs. others, MED vs. others, RHAB vs. others and GLIO vs. others) were defined. A number of classifiers were trained on probe set subsets of increasing size (n = 10, 20, 50, 100, 200, 500, 1000). Probe sets were selected here using a signal-to-noise ratio (SNR) variation filter, i.e. $|\mu_1-\mu_2| \,/\, \sqrt{(\sigma_1^2 + \sigma_2^2)}$ on the training set. Classifiers used were nearest centroid (NC), nearest shrunken centroid (PAM) (30), LIKNON (31), *k*-nearest neighbour (*k*-NN), support vector classifier with polynomial kernel of degree d (SVC-P) and radial basis function kernel of width σ (SVC-R) (32). The parameters k, d and σ were optimised by performing cross-validation (k: leave-one-out; d, σ: 10-fold) on the training set only. Both PAM and LIKNON provide their own feature selection algorithm, which selects

the optimal feature set within the set selected by the variation filter. In a single experiment, 90 percent of the samples (randomly selected) were used to train a classifier after which the classifier was tested on the remaining 10 percent. This experiment was repeated 100 times, resulting in an average performance and a standard deviation.

3.4 Results and discussion

The aim of our work was to evaluate the effect of several microarray data pre-processing methods on the outcome of analyses commonly applied in patient-cohort studies. Four types of analysis were performed: (1) expression level measurement, (2) detection of differential expression (supervised), (3) cluster analysis (unsupervised) and (4) classification of samples (supervised). These analyses were applied to two publicly available datasets, one of 285 acute myeloid leukemia (AML) samples profiled on Affymetrix HG-U133A GeneChips (17) and one consisting of 42 Affymetrix HuGeneFL GeneChips hybridized with central nervous system (CNS) tumor tissue (20).

3.4.1 Comparing expression levels to RQ-PCR (AML dataset)

Pearson correlation coefficients between expression levels of *EVI1*, *CEBPa*, *MEIS1*, *HOXA7*, *HOXA9*, *TRKA*, *PRDM1*, *PRDM2*, *P8* and *GMCSF* found in the AML dataset, measured by RQ-PCR and microarray after pre-processing using different methods, are listed in Table 2. The actual expression levels measured by RQ-PCR and the Affymetrix probe sets (using the different pre-processing methods) are listed in Supp. Table 2. Average correlation to RQ-PCR expression is 0.48-0.57 for the different methods with RMA (0.57 ± 0. 30) and GCRMA (0.57 ± 0.28) showing the highest correlation on average, dChip (0.48 ± 0.31) scoring lowest and MAS (0.52 ± 0.29) scoring intermediate. No significant differences have been found between correlations of different pre-processing methods and RQ-PCR data and (taking RQ-PCR as gold standard) no pre-processing method unequivocally performed best in measuring expression level. Correlation overall is moderate, but this result is likely to be influenced by different genomic location of RQ-PCR primers and Affymetrix probes, resulting in different expression values when alternative splicing occurs; by incorrect annotation of individual probe sets (such as 206848_at); and by suboptimal RQ-PCR primer and Affymetrix probe design.

3.4.2 Comparing expression levels between pre-processing methods

Correlations are depicted in Figure 1A for the AML dataset and in Figure 1B for the CNS dataset for each probe set present on the microarray, ordered by average expression level over the four differently pre-processed datasets. Overall, a clear trend of increasing correlation at increasing expression levels is apparent, which has been noticed before (15). Aside from a dense area of highly correlated genes with intermediate to high expression, in several comparisons, for instance that of RMA to MAS, a second more densely populated area is visible in the range of extremely low expression levels. These expression levels correspond to non-expressed genes (40-50% of all probe sets). At these levels, variability is relatively higher, resulting in moderate correlations. As the normalization method of RMA

Gene symbol	Probe set ID	MAS	dChip	RMA	GCRMA
EVI1	215851_at	0.34	0.52	0.63	0.29
	221884_at	0.64	0.87	0.88	0.45
P8	209230_s_at	0.09	0.15	0.16	0.00
PRDM1	217192_s_at	0.53	0.57	0.55	0.53
TRKA	208605_s_at	0.66	0.77	0.75	0.63
PRDM2	205277_at	0.21	0.25	0.28	0.25
	203057_s_at	0.56	0.56	0.59	0.58
	203056_s_at	0.57	0.59	0.59	0.55
	216433_s_at	0.33	0.42	0.45	0.12
MEIS1	204069_at	0.88	0.90	0.90	0.86
HOXA9	214651_s_at	0.90	0.90	0.91	0.89
	209905_at	0.91	0.91	0.90	0.90
HOXA7	206847_s_at	0.80	0.77	0.80	0.89
	206848_at	0.01	0.02	0.02	0.04
CEBPa	204039_at	0.61	0.61	0.63	0.58
GMCSF	210229_s_at	0.20	0.32	0.10	0.15

Table 2. Correlation between expression levels measured by RQ-PCR and Affymetrix GeneChip on the AML dataset, after use of different pre-processing methods.

and GCRMA is the same (both using MM probes only for background correction) and their summarization methods are very similar, it is not surprising that these methods show the highest resemblance in measured expression. However, they show more agreement with MAS than with dChip (which was also seen when comparing microarray expression levels to RQ-PCR expression levels). Perhaps this has to do with the fact that dChip calculates expression on the original probe intensity values rather than the log-transformed ones used by the other methods.
The CNS dataset shows similar trends, but the much higher level of variation suggests that sample size and/or quality of the platform and biological sample have a much more profound effect on estimated expression levels, than has the pre-processing method.
In conclusion, the AML dataset shows that variation in estimated expression levels exists between different pre-processing methods and that this variation is higher at lower mRNA concentrations. The clear trend of increasing correlation with increasing expression level suggests that pre-processing has an influence, but this concerns only a minority of probe sets.
Using the Affymetrix Latin square dataset, Rajagopalan noted that MAS and dChip perform equally well on estimating expression levels, with a small non-significant

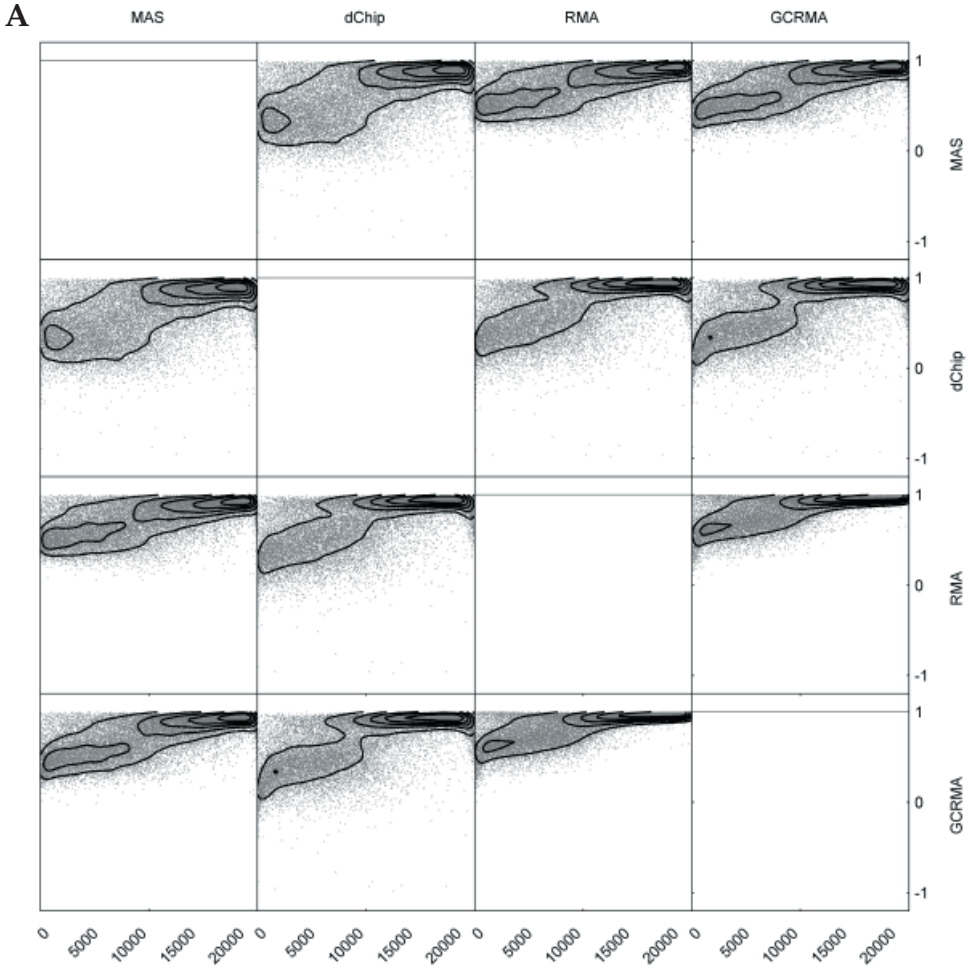
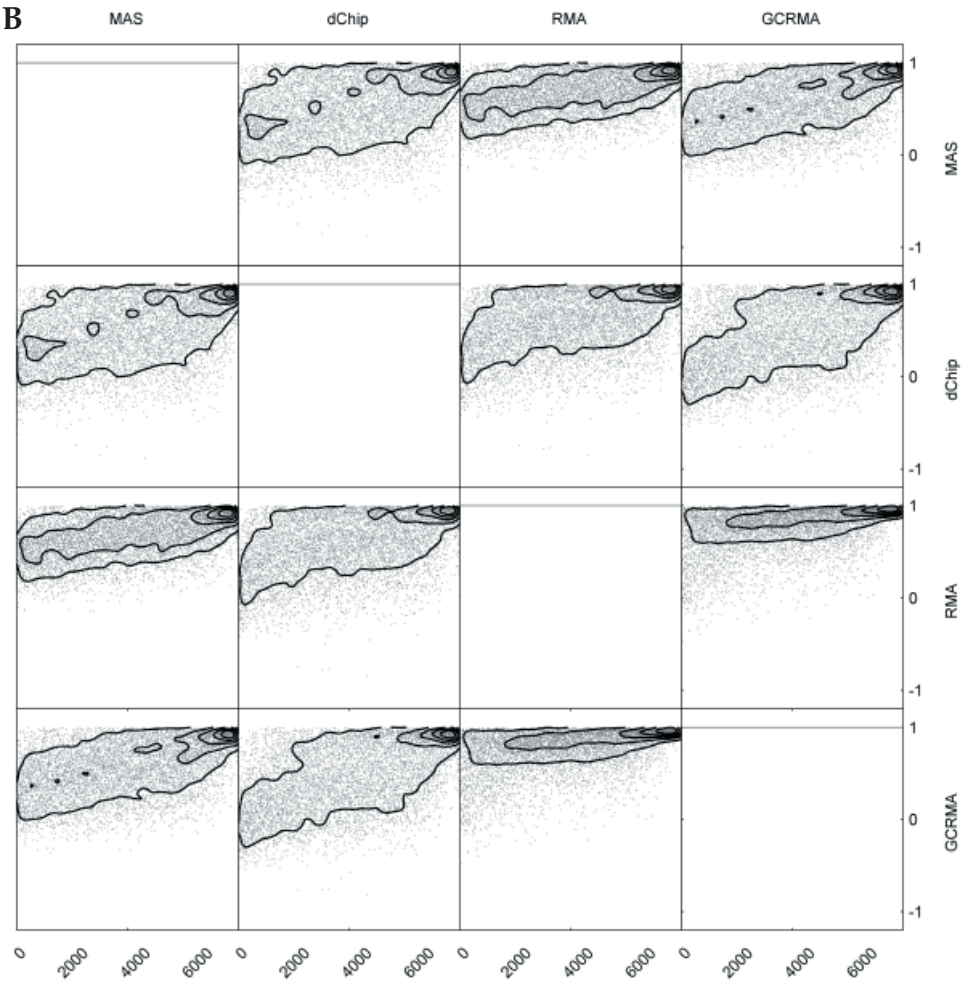


Figure 1: Correlation of expression values pre-processed by two methods. *Pearson correlation coefficients of expression measurements calculated by two pre-processing procedures are shown on the y-axis, probe sets ranked by average expression level over the four pre-processing methods are shown on the x-axis. Contours indicate equal density, as estimated using a Gaussian kernel density estimate, with kernel width optimised by leave-one-out maximum-likelihood. A. AML dataset. B. CNS dataset.*

advantage for MAS (13). Although trends towards MAS seem visible in our study as well, MAS and dChip behave rather differently in the experimental dataset used here.

Testing not only the Latin square dataset but also the GeneLogic dataset, Irizarry et al. conclude that RMA shows highest sensitivity and specificity when compared to dChip and the AvDiff algorithm (10). As no method performs significantly different in our study, these results are not confirmed.



3.4.3 Differential expression

Significance of differences in expression when comparing two conditions was calculated using three standard methods: the *t*-test; the Wilcoxon rank sum test, controlling the Family-Wise Error Rate (FWER); and Significance Analysis of Microarrays (SAM), a test controlling the False Discovery Rate (FDR) using a statistic resembling that of the *t*-test (28). Different pre-processing methods were compared by assessing the overlap in the number of probe sets marked as differentially expressed by two pre-processing methods. We call a probe set differentially expressed below an FWER or FDR of 5%. In the AML-dataset, *p*-values (FWER) and *q*-values (FDR) were computed for samples with recurrent *FLT3* ITD mutations vs. the rest, inv(16) vs. the rest, t(15;17) vs. the rest and t(8;21) vs. the rest. In the CNS-dataset, *p*-values and *q*-values were computed for PNET-, RHAB-, GLIO- and MED-samples vs. the rest, respectively. Although different subdivisions into conditions were thus compared, the outcomes are remarkably similar.

Considering the AML dataset, the overlap between the probe sets selected on RMA and GCRMA pre-processed data is most striking, with a minimum *R* of 0.78 (average 0.85, Table 3A, Supp. Tables 3A-C). Overall, the overlap between different pre-processing methods is considerable: a minimum *R* of 0.56 (average 0.74) is found, independent of the statistical test used. The combination MAS-dChip comes up as least comparable (Table 3A, Supp. Tables 3A-C). MAS shows higher concordance with RMA and GCRMA than does dChip. No indications were found that *R* will increase for smaller FWER or FDR (data not shown).

The overlap between probe set lists detected as differentially expressed is considerably less in the CNS dataset than in the AML dataset and there is more variation, which could be due to the higher amount of noise in this dataset and/or its smaller sample size. The RMA-GCRMA comparison results in an average *R* of 0.56 (Table 3B, Supp. Tables 3D-F). Again, pre-processing with MAS will result in less differences with RMA pre-processing than with dChip (Table 3B, Supp. Tables 3D-F). Overall, average *R* is 0.40 for this dataset. When using *q*-values, again RMA and/or MAS often detect larger numbers of differentially expressed probe sets than dChip and GCRMA. Note also that the difference between the number of probe sets selected using the *t*-statistic and the Wilcoxon statistic is larger than for the AML dataset. This may be caused by outlier data on the HuGeneFL microarrays, to which the *t*-test is more susceptible.

In a study evaluating experimental datasets of 79 ovary tumors and 47 colon tumors profiled on the Affymetrix HG-U133A platform, Shedden et al. (9) show that dChip results are closer to those obtained using RMA than to those obtained using MAS, an observation not confirmed by our results. Statistical tests on RMA and MAS pre-processed data detect the largest number of differentially expressed probe sets in most cases, where GCRMA and dChip select less, with a maximum difference in number of selected probe sets of 49.7% in the AML dataset (Supp. Table 3A, *q*-values). This does not confirm the observations of Shedden et al., who found that dChip outperformed MAS and GCRMA in terms of sensitivity. Irizarry et al. (10) report that RMA performs better than dChip and the AvDiff algorithm in finding truly differentially expressed genes. No statement on the true nature of probe sets measured as differentially expressed here can be made. However, MAS and RMA score roughly equal numbers of probe sets as differentially expressed and both methods find more probe sets to be differentially expressed than dChip and GCRMA, as in (10).

Recently, Hoffmann et al. (11) stated that normalization will have a larger influence on the number of differentially expressed genes than the actual statistical test used. Although a direct comparison of (11) and our work is not possible due to differences in multiple testing correction, in our (much) larger datasets we observe a larger difference between the number of probe sets selected as a result of the multiple testing correction used (FWER or FDR) than as a result of the choice of pre-processing method.

Overall, the overlap between sets of genes selected as differentially expressed is considerable when pre-processing the data using different methods and overlap increases when non-biological variation decreases. Using the current datasets, it is not possible to give indications of the quality of probe sets selected, due to the lack of ground truth.

		MAS	dChip	RMA	GCRMA
SAM *q*-values	MAS	3185			
	dChip	0.68	2973		
	RMA	0.72	0.72	3649	
	GCRMA	0.73	0.70	0.86	3650
t-test *p*-values	MAS	458			
	dChip	0.66	354		
	RMA	0.68	0.70	419	
	GCRMA	0.69	0.71	0.83	472
Wilcoxon test *p*-values	MAS	337			
	dChip	0.72	295		
	RMA	0.75	0.79	322	
	GCRMA	0.75	0.79	0.87	344

Table 3A. Overlap R(A,B) between sets of genes marked as differentially expressed after pre-processing with different methods. *p- and q-values for the significance of difference in expression between samples from the AML dataset with recurrent FLT3 mutation and samples without recurrent FLT3 mutation were calculated. The numbers on the diagonal represents the number of probe sets marked as differentially expressed after application of each method.*

		MAS	dChip	RMA	GCRMA
SAM *q*-values	MAS	400			
	dChip	0.39	714		
	RMA	0.50	0.54	666	
	GCRMA	0.52	0.39	0.58	330
t-test *p*-values	MAS	224			
	dChip	0.36	303		
	RMA	0.43	0.51	159	
	GCRMA	0.47	0.37	0.55	123
Wilcoxon test *p*-values	MAS	17			
	dChip	0.41	17		
	RMA	0.46	0.29	18	
	GCRMA	0.45	0.26	0.5	14

Table 3B. CNS dataset, GLIO samples vs. others.

3.4.4 Cluster analysis

Data resulting from different pre-processing methods was clustered by *k*-means (KM) and hierarchical clustering with single, average and complete linkage (HC/S, HC/A, HC/C). Clusterings of both the AML and CNS datasets were compared using the Jaccard index; results are shown in Figure 2 and Supp. Figure 1 (33). RMA and GCRMA results are often similar, which is to be expected. dChip results frequently differ from results obtained using other pre-processing methods. In general, KM, HC/A and HC/C on both datasets show Jaccard indices of 0.3-0.6. HC/S shows higher indices, but the actual resulting clusterings are very poor due to the well-known high susceptibility of this method to outliers (data not shown): almost all samples end up in a single cluster, the remaining samples form individual clusters. As an example, the confusion matrix in Table 4 shows that many clusters found using the MAS pre-processed dataset are also found reasonably well using the RMA pre-processed dataset (by *k*-means clustering into $k = 12$ clusters, on correlation distance, using 3000 probe sets). However, as there are 2716 sample pairs co-occurring in a cluster in both clustering results, 1099 sample pairs co-occurring in a cluster in the MAS clustering result only and 1137 sample pairs co-occurring in a cluster in the RMA result only, this leads to a Jaccard-index *J* of only $2716 / (2716 + 1099 + 1137) = 0.55$.

In an attempt to quantify the sensitivity of clusterings found to small perturbations, stability-normalized Jaccard indices J^{SN} were therefore calculated, indicating to what extent the Jaccard indices *J* found are out of the ordinary. Figure 3A illustrates

		Number of samples in RMA clusters											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Number of samples in MAS clusters	**1**	33	1	1	0	1	1	1	0	0	1	1	0
	2	0	33	0	0	1	0	0	0	0	0	2	0
	3	0	3	28	2	0	0	0	0	0	0	0	0
	4	0	0	7	24	0	1	0	1	0	0	0	0
	5	0	1	0	0	21	2	0	0	0	0	0	0
	6	0	0	1	3	0	21	0	0	0	0	0	0
	7	0	0	0	0	0	0	21	0	0	0	0	0
	8	0	0	0	1	0	0	0	18	0	0	0	0
	9	0	0	0	0	4	0	0	0	10	0	0	8
	10	0	0	0	0	0	0	0	0	1	10	0	0
	11	0	3	0	0	0	0	0	0	0	0	10	0
	12	0	0	1	0	0	0	0	0	7	0	0	0

Table 4. Confusion matrix of MAS and RMA clustering results. *Clustering into $k = 12$ clusters was performed using k-means clustering, using correlation distance on 3000 probe sets. A cell at position (i,j) shows the number of samples assigned to cluster i on data pre-processed using MAS and to cluster j on data pre-processed using RMA. The Jaccard index between these two clusterings is 0.55.*

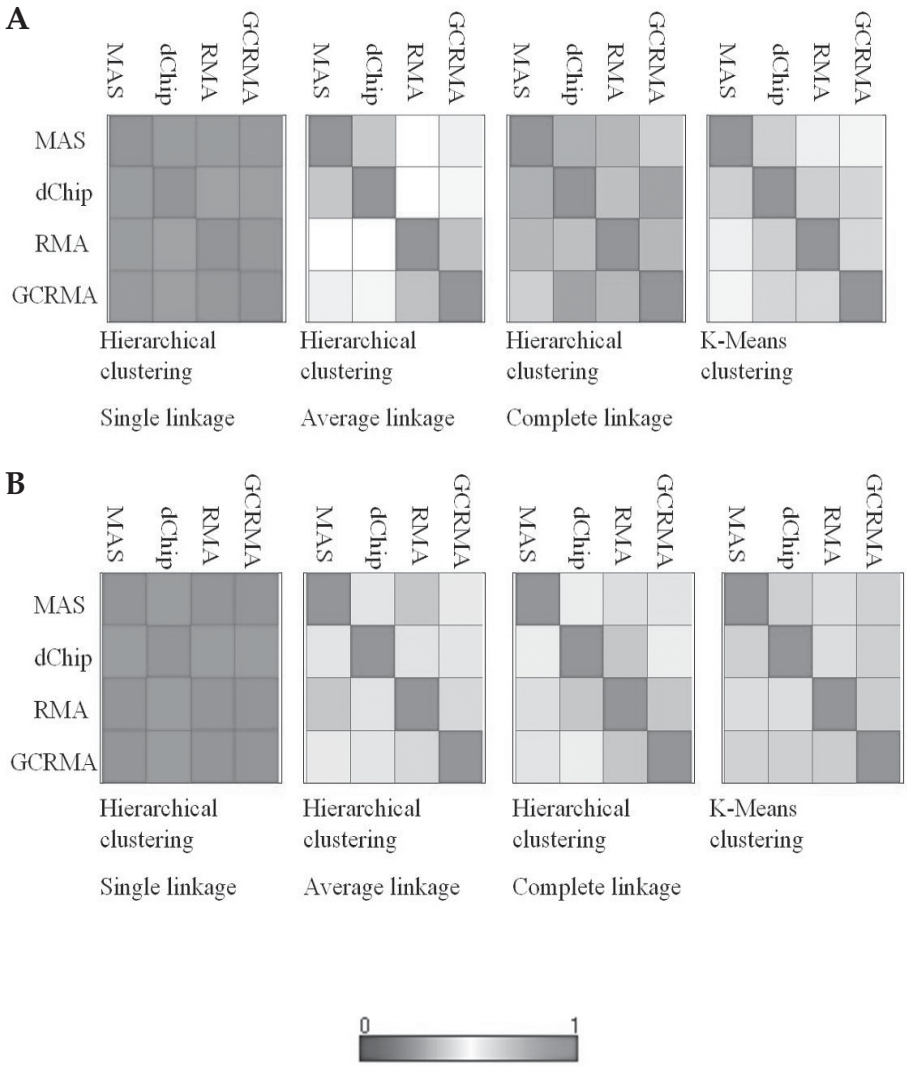


Figure 2: Jaccard indices of clustering results. *Results were obtained using correlation distance on a fixed number of probe sets, after different pre-processing procedures and by different clustering algorithms. A. AML dataset, k = 12 clusters, 3000 probe sets. B. CNS dataset, k = 5 clusters, 1000 probe sets.*

A full-color version of this figure is provided on the CD.

that for KM and the pair of pre-processing methods used (MAS and RMA), $J = 0.55$ is actually better than the Jaccard index obtained on average on a slightly changed version of the MAS pre-processed dataset ($J^{SN} > 0.5$), but worse than that obtained on average on a slightly changed version of the RMA pre-processed dataset ($J^{SN} < 0.5$).

Figure 3B shows that for KM and HC/A, differences using MAS and (GC)RMA are actually roughly of the same order as differences between 90% subsamples of the MAS pre-processed dataset (i.e. the J^{SN} is high for MAS). To a lesser extent, this also

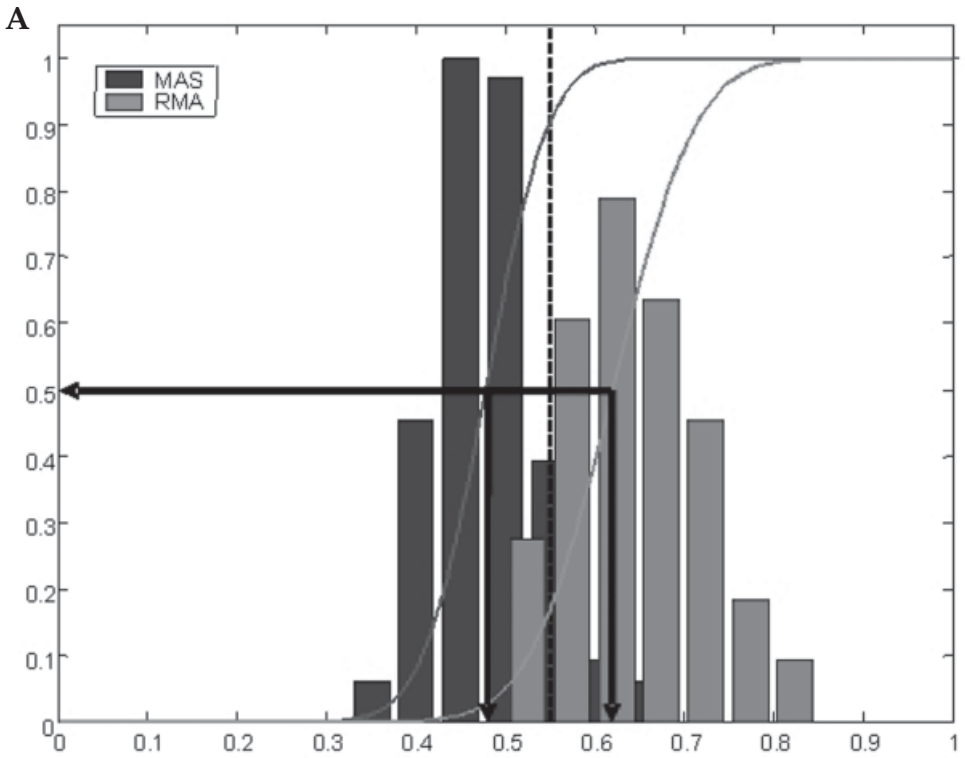


Figure 3A: Stability normalization of Jaccard index. *Illustration of stability normalization for the Jaccard index of a particular k-means clustering (k = 12), obtained on MAS- and RMA-pre-processed versions of the AML dataset (correlation distance, 3000 probesets). The dotted line corresponds to the Jaccard index between these clusterings (0.55). For both MAS and RMA, the CDF can be used to arrive at a stability normalized Jaccard index; in this case 0.90 and 0.16. The arrows indicate the Jaccard indices for which the normalised Jaccard index J^{SN} = 0.5. The interpretation is that for MAS, the comparison to RMA falls well within what can be expected, for RMA less so.*

A full-color version of this figure is provided on the CD.

holds for dChip vs. (GC)RMA. However, these same differences are quite large in terms of the differences in clusterings between 90% subsamples of (GC)RMA (i.e. the J^{SN} is low for (GC)RMA). The main cause for this is (GC)RMA's higher stability: as it normalises over all arrays – unlike MAS and dChip – leaving out a small subset will have only a limited effect on probe set distributions, and hence on clustering results. When RMA and GCRMA results are compared to each other, a high J^{SN} results as well. HC/C oftens shows lower values for J and J^{SN}.

The CNS dataset (Figure 3C) largely tells the same story, although the J^{SN} are somewhat larger, especially for k-means clustering. This is due to the smaller sample size: leaving out 10% of the samples relatively has more impact on the Jaccard indices.

Supp. Figures 1 and 2 illustrate the influence of the choice of the number of clusters (k) and the distance measure (correlation or Euclidean). Both datasets show the

B

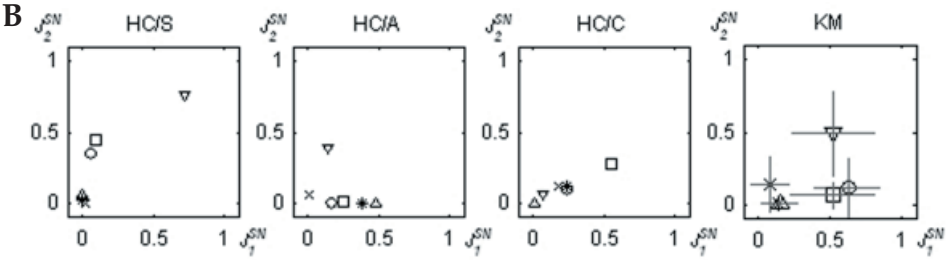


Figure 3B. AML dataset: stability-normalized pairwise Jaccard indices of cluster labels assigned by the various methods. *Clusterings into k = 12 clusters obtained using correlation distance on 3000 probe sets. Legend is shown in Figure 3D. For k-means, the grey bars indicate standard deviation over 10 repeated experiments.*

C

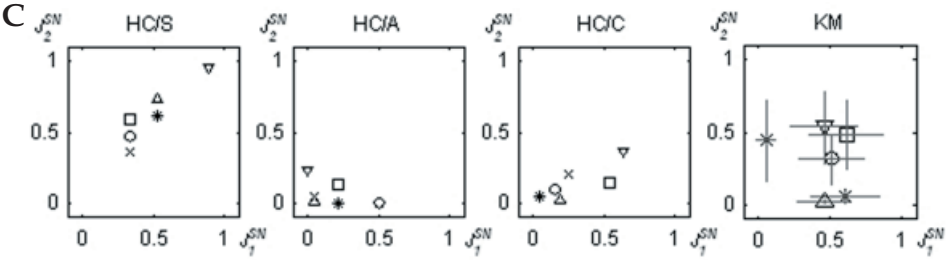


Figure 3C. CNS dataset: stability-normalized pairwise Jaccard indices of cluster labels assigned by the various methods. *Clusterings into k = 5 clusters obtained using correlation distance on 1000 probe sets. Legend is shown in Figure 3D. For k-means, the grey bars indicate standard deviation over 10 repeated experiments.*

D

Figure 3D. Legends to markers in Figures 3B-C.

same effects. For lower k (k = 2), both Jaccard indices and stability-normalized Jaccard indices are much higher, as clusterings of data pre-processed by the various methods agree on structure clearly present in the data. For higher k (AML: k = 20, CNS: k = 10), Jaccard indices and stability-normalized Jaccard indices are similar to or even lower than those for the k chosen originally. Using Euclidean distance leads to slightly lower Jaccard indices, with an increase in difference between dChip and other methods. This may be the result of the negative values it produces (unlike MAS and (GC)RMA), which are thresholded at 0.1 in the data transformation steps. Due to the centering by the geometric mean this can lead to larger extreme probe set values over arrays.

In conclusion, clustering results are sensitive to the choice of pre-processing method.

This sensitivity is smallest for small numbers of clusters k (i.e. when looking for clearly present structure) and when using correlation distance. Additionally, using (GC)RMA seems to result in more stable clusterings than using MAS or dChip.

3.4.5 Classification

A number of different classification problems defined on the datasets have been approached using several classifiers trained on data of all pre-processing methods. Resulting performances are listed in Table 5 and Supp. Table 4. Results are reported only for the number of probe sets giving lowest average test set error over the four methods. Although this makes the performance estimates biased, it does not influence comparison between methods.

In the AML dataset, inversion of chromosome 16 is well predictable, with error rates smaller than 5% (Table 5a). Differences in error rate between classification algorithms are very small: although the nearest centroid classifier often performs worst, no algorithm performs significantly better than others. More importantly, no pre-processing method scores significantly better or worse than others (although MAS relatively often shows best results). Although predicted with a higher error rate, these observations are confirmed on the *FLT3* (Table 5B) and CCR (Table 5C) AML problems.

Interestingly, this also holds for the CNS dataset (Table 5D shows results for the MED problem; other results are shown in Supp. Table 4), although performances show much more variation and MAS no longer comes out best. Ofcourse, the CNS dataset is rather small, so obtaining good classifiers is harder.

No classifier or pre-processing method scores significantly better than others. This can be explained by the fact that the probe sets on which classification is based are already selected to give good classification results: on differently pre-processed datasets, different probe sets may be selected (in fact, the selected sets of n = 1000 probe sets show an overlap of 71% to 87%). The six classifiers used seem to be equally susceptible to different pre-processing methods; that is, for each of them performance varies with pre-processing method used in at least some of the problems. For classification, the choice of pre-processing method (and, for that matter, classification algorithm) seems to be irrelevant.

3.5 Conclusion

Patient-cohort studies using microarrays are often performed to find pathobiologically relevant relations between genes and patient classes. The Affymetrix platform has become increasingly popular for this type of study. Processing intensity values obtained using Affymetrix GeneChips remains a challenging task for many microarray researchers. Apart from the Affymetrix MAS procedure, several statistical procedures have been proposed to assess expression, such as dChip, RMA and GCRMA. Our study has tried to estimate the effects of the choice of pre-processing method from a practical viewpoint. To this end, we have applied a number of analyses to two datasets, which we believe to represent two extremes in recent patient-cohort studies both in terms of sample size and of platform used.

The experimental results indicate that the normalization step in (GC)RMA has a larger effect on the data than the one in MAS and dChip, but this cannot be

Classifier (number of probe sets used)	MAS	dChip	RMA	GCRMA
NC (50)	0.01(0.02)	0.02(0.02)	0.01(0.02)	0.02(0.03)
PAM (20)	0.01(0.02)	0.01(0.02)	0.01(0.02)	0.01(0.02)
LIKNON (10)	0.02(0.03)	0.02(0.03)	0.02(0.02)	0.02(0.02)
k-NN (50)	0.01(0.02)	0.01(0.02)	0.01(0.02)	0.01(0.02)
SVC/P (10)	0.02(0.03)	0.02(0.02)	0.02(0.03)	0.02(0.03)
SVC/RBF (10)	0.03(0.02)	0.02(0.02)	0.02(0.02)	0.02(0.02)

Table 5A. Performance of different classification algorithms: AML dataset, inv(16) problem. *Mean test set error (standard deviation) over 100 random splits of the original data into a training set (90%) and a test set (10%). Error is defined as average error per class, i.e. corresponding to assuming a prior probability of occurrence of a class of 50%. Classifiers were trained for 10, 20, 50, 100, 250, 500 and 1000 probe sets selected by the variation filter; results shown here are for the number of probe sets resulting in the smallest average test set error over the four methods, indicated between brackets after the classifier name.*

Classifier (number of probe sets used)	MAS	dChip	RMA	GCRMA
NC (10)	0.16(0.06)	0.19(0.09)	0.16(0.08)	0.26(0.08)
PAM (20)	0.14(0.06)	0.14(0.06)	0.14(0.06)	0.14(0.06)
LIKNON (20)	0.12(0.05)	0.11(0.05)	0.12(0.06)	0.13(0.05)
k-NN (200)	0.10(0.05)	0.11(0.05)	0.11(0.06)	0.13(0.06)
SVC/P (20)	0.12(0.05)	0.11(0.05)	0.13(0.06)	0.13(0.05)
SVC/RBF (100)	0.09(0.05)	0.10(0.05)	0.10(0.05)	0.14(0.05)

Table 5B. AML dataset, FLT3 problem.

Classifier (number of probe sets used)	MAS	dChip	RMA	GCRMA
NC (1000)	0.34(0.09)	0.34(0.08)	0.33(0.08)	0.35(0.08)
PAM (10)	0.29(0.06)	0.30(0.06)	0.30(0.05)	0.30(0.06)
LIKNON (20)	0.27(0.08)	0.31(0.07)	0.31(0.08)	0.28(0.07)
k-NN (20)	0.28(0.07)	0.29(0.06)	0.30(0.06)	0.29(0.07)
SVC/P (20)	0.28(0.08)	0.31(0.07)	0.31(0.07)	0.28(0.06)
SVC/RBF (20)	0.28(0.05)	0.30(0.04)	0.30(0.04)	0.29(0.04)

Table 5C. AML dataset, CCR problem.

Classifier (number of probe sets used)	MAS	dChip	RMA	GCRMA
NC (1000)	0.04(0.09)	0.03(0.09)	0.03(0.08)	0.07(0.13)
PAM (1000)	0.05(0.10)	0.07(0.13)	0.09(0.14)	0.06(0.13)
LIKNON (1000)	0.10(0.12)	0.06(0.11)	0.10(0.13)	0.18(0.18)
k-NN (500)	0.06(0.11)	0.06(0.12)	0.08(0.12)	0.04(0.10)
SVC/P (1000)	0.09(0.12)	0.05(0.11)	0.04(0.09)	0.07(0.12)
SVC/RBF (10)	0.17(0.13)	0.16(0.13)	0.17(0.14)	0.14(0.14)

Table 5D. CNS dataset, MED problem.

separated from the effect of applying different models for summarization. And, although the dChip and RMA summarization models are more related to each other than the MAS and RMA ones, MAS pre-processed data shows more similarity to RMA than does dChip.

In practical terms, the question of which method will give expression value estimates closest to the actual data is still to be answered; this study has not attempted to answer it, because we have not used data with accompanying ground truth. We showed that results of various analyses are not always dependent on the choice of pre-processing method. Analyses such as calculating expression levels or

assessing differential expression are reasonably susceptible to differences between pre-processing methods; clustering as well, except when looking for clearly present structure (that is, using a small number of clusters); but classification far less so. The message is that while care should be taken in assigning biological meaning to individual probe set measurements, this holds less for global statements about the data.

Several other studies have been performed to assess the level of concordance in differential gene sets between pre-processing methods and noted that the choice of the method was of major influence, with different studies favoring different pre-processing methods (9, 11, 13). Our results do not conclusively confirm one or more studies, although results partially overlap. One major difference with other studies is the size of the used datasets, where one of the datasets used in this study is considerably larger. It is to be expected that with the evolution of the array technology, the number of profiled samples in any single patient-cohort study is likely to increase.

The effects of the choice of pre-processing method are far more profound in the CNS dataset than in the AML dataset. Several possible explanations can be given for this, but it is not possible to single any of them out based only on the two datasets used in this study. The AML dataset contains more samples, which allows for better parameter estimates in the analysis methods presented in this work. Furthermore, Affymetrix technology has evolved over time, resulting in a more stable platform for the AML dataset (HG-U133A) than the CNS dataset (HuGeneFL). Biological differences also play a role in the two datasets. The amount of viable cells obtained from bone marrow is also likely to be higher compared to solid tumors, which often show necrotic areas, leading to difference in RNA-quality and -degradation. Also, tumor cells can be purified from bone marrow samples using Ficoll-centrifugation, a technique which is not available for the solid tumors which were hybridized in the CNS dataset, resulting in less contamination with other cell types in hybridized samples, which is known to be an important factor (34). We recommend that the emphasis in setting up a large microarray-based study should therefore be on the quality of the biological sample and the quality of RNA rather than on the choice of the pre-processing procedure. However, we do believe that an inverse relation exists, with the importance of the method of normalization and expression summarization increasing when the quality of the biological sample and the number of studied samples decrease. Although we base this on a limited number of pre-processing methods and data sets, we think that taking into account more available methods will have no effect on our conclusion.

List of Abbreviations

AML	Acute myeloid leukemia
CNS	Central nervous system
PNET	Primitive neuro-ectodermal tumors
MED	Medullablastoma
GLIO	Malignant glioma
RHAB	Rhabdoid tumors
SAM	Significance analysis of microarrays
CDF	Cumulative distribution function

CCR	Continuous complete remission
PAM	Prediction analysis of microarrays
k-NN	k-Nearest neighbour
NC	Nearest centroid
SVC-P	Support vector classifier with polynomial kernel of degree d
SVC-R	Support vector classifier with radial basis function kernel of width σ

Acknowledgements

The authors would like to thank Sahar Barjesteh van Waalwijk van Doorn – Khosrovani, Renee Beekman, Claudia Erpelinck , Judith Gits and Antoinette Van Hoven – Beijen for providing RQ-PCR data and Ruud Delwel for helpful discussions. The authors are grateful to the anonymous reviewers for their helpful remarks, which improved the manuscript.

References

1. Affymetrix: Microarray Suite User Guide. In., vol. Version 5; 2001.
2. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 2001, 98(1):31-36.
3. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003, 19(2):185-193.
4. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003, 31(4):e15.
5. Naef F, Magnasco MO. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Phys Rev E Stat Nonlin Soft Matter Phys 2003, 68(1 Pt 1):011906.
6. Wu Z, Irizarry RA, Gentleman R, Murillo F, Spencer F. A model based background adjustment for oligonucleotide expression arrays. In. Baltimore: John Hopkins University; 2004.
7. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002, 18 Suppl 1:S96-104.
8. Affymetrix: Guide to probe logarithmic intensity error (PLIER) estimation. In.; 2005.
9. Shedden K, Chen W, Kuick R, et al. Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. BMC Bioinformatics 2005, 6(1):26.
10. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003, 4(2):249-264.
11. Hoffmann R, Seidl T, Dugas M. Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. Genome Biol 2002, 3(7): RESEARCH0033.
12. Liu WM, Mei R, Di X, et al. Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics 2002, 18(12):1593-1599.
13. Rajagopalan D. A comparison of statistical methods for analysis of high density oligonucleotide array data. Bioinformatics 2003, 19(12):1469-1476.
14. Freudenberg J, Boriss H, Hasenclever D. Comparison of preprocessing procedures for oligonucleotide micro-arrays by parametric bootstrap simulation of spike-in experiments. Methods Inf Med 2004, 43(5):434-438.
15. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 2004, 20(3):323-331.
16. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 2004, 350(16):1605-1616.

17. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 2004, 350(16):1617-1628.
18. Yeoh EJ, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 2002, 1(2):133-143.
19. Su AI, Welsh JB, Sapinoso LM, et al. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001, 61(20):7388-7393.
20. Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002, 415(6870):436-442.
21. Van der Reijden BA, de Wit L, van der Poel S, et al. Identification of a novel CBFB-MYH11 transcript: implications for RQ-PCR diagnosis. Hematol J 2001, 2(3):206-209.
22. Barjesteh van Waalwijk van Doorn-Khosrovani S, Erpelinck C, van Putten WL, et al. High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. Blood 2003, 101(3):837-845.
23. van Waalwijk van Doorn-Khosrovani SB, Erpelinck C, Meijer J, van Oosterhoud S, 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.
24. Mulloy JC, Jankovic V, Wunderlich M, et al. AML1-ETO fusion protein up-regulates TRKA mRNA expression in human CD34+ cells, allowing nerve growth factor-induced expansion. Proc Natl Acad Sci U S A 2005, 102(11):4016-4021.
25. Bengtsson M, Stahlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Res 2005, 15(10):1388-1392.
26. Sheskin DJ. Handbook of parametric and nonparametric statistical procedures., Third edition edn. Boca Raton, FL: Chapman & Hall/CRC; 2004.
27. Ge U, Dudoit S, Speed TP. Resampling-based multiple testing for microarray analysis. In.; 2003.
28. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001, 98(9).5116-5121.
29. Lange T, Roth V, Braun ML, Buhmann JM. Stability-based validation of clustering solutions. Neural Comput 2004, 16(6):1299-1323.
30. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002, 99(10):6567-6572.
31. Bhattacharyya C, Grate LR, Rizki A, et al. Simultaneous relevant feature identification and classification in high-dimensional spaces: Application to molecular profiling data. Signal Processing 2003, 83:729-743.
32. Duda RO, Hart PE, Stork DG. Pattern classification, second edition edn. Hoboken, NY: Wiley Interscience; 2003.
33. Jain KJ, Dubes RC. Algorithms for clustering data. Upper Saddle River, NJ: Prentice Hall Inc.; 1988.
34. de Ridder D, van der Linden CE, Schonewille T, et al. Purity for clarity: the need for purification of tumor cells in DNA microarray studies. Leukemia 2005, 19(4):618-627.

Chapter 4

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

Roel G.W. Verhaak[1], Chantal S. Goudswaard[1], Wim van Putten[2], Maarten A. Bijl[1], Mathijs A. Sanders[1], Wendy Hugens[3], André G. Uitterlinden[3], Claudia A.J. Erpelinck[1], Ruud Delwel[1], Bob Löwenberg[1], and Peter J.M. Valk[1]

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

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

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

Blood. 2005 Dec 1;106(12):3747-54.

4.1 Abstract

Mutations in nucleophosmin *NPM1* are the most frequent acquired molecular abnormalities in acute myeloid leukemia (AML). We determined the *NPM1* mutation status in a clinically and molecularly well-characterized patient cohort of 275 patients with newly diagnosed AML by denaturing high performance liquid chromatography (dHPLC). We show that *NPM1* mutations are significantly underrepresented in patients younger than 35yr. *NPM1* mutations positively correlate with AML with high white blood cell counts, normal karyotypes and *FLT3* internal tandem duplication (ITD) mutations. *NPM1* mutations associate inversely with the occurrence of *CEBPa*- and N-*RAS* mutations. With respect to gene expression profiling, we show that AML cases with an *NPM1* mutation cluster in specific subtypes of AML with previously established gene expression signatures, are highly associated with a homeobox gene-specific expression signature and can be predicted with high accuracy. We demonstrate that patients with intermediate cytogenetic risk AML without *FLT3* ITD mutations but with *NPM1* mutations have a significantly better OS and EFS than those without *NPM1* mutations. Finally, in multivariable analysis *NPM1* mutations express independent favorable prognostic value with regard to overall-, event free- and disease free survival.

4.2 Introduction

Acute myeloid leukemia (AML) is a heterogeneous disease with diverse genetic abnormalities and variable responsiveness to therapy. Cytogenetic analyses and molecular analyses are currently used to risk-stratify AML. For instance, the translocations inv(16), t(8;21), and t(15;17) herald a favorable prognosis, whereas certain other cytogenetic aberrations indicate leukemia with intermediate or high risk of relapse (1-5). Nevertheless, the classification of AML on the basis of karyotyping is still far from satisfactory. In recent years extended molecular analyses have yielded novel molecular markers important for proper diagnostics of AML. The ITD in the fms-like tyrosine kinase-3 gene (*FLT3*) (6, 7), partial tandem duplication (PTD) (8, 9) of the mixed lineage leukemia gene (*MLL*) and increased expression of the transcription factor EVI1 (10), are indicative of poor prognosis. In contrast, mutations in the transcription factor *CEBPa* have been associated with a favorable response to therapy (11, 12). A recent study showed mutations in exon 12 of the gene encoding nucleophosmin *NPM1* in approximately 35% of cases of *de novo* AML (13). Mutations of *NPM1* were found to be mutually exclusive with certain common recurrent chromosomal aberrations, and are predominantly seen in AML with normal karyotypes and *FLT3* ITD mutations.

NPM1 is predominantly localized in the nucleolus and is thought to function as a molecular chaperone of proteins, facilitating the transport of ribosomal proteins through the nuclear membrane (14-16). Disruption of *NPM1*, either by chromosomal translocation or by mutation, results in the cytoplasmic dislocation of *NPM1*. The high frequency of *NPM1* mutations in AML with normal karyotypes and the observation that cytoplasmic NPM1 cannot exert its normal functions as binding partner and transporter protein lead to the notion that *NPM1* mutation may be an early event in leukemogenesis.

An important role for *NPM1* in leukemias and lymphomas has been proposed previously. *NPM1* has been found to be part of several fusion proteins, which are

formed as a result of chromosomal translocation and in which only the NPM1 N-terminal region is conserved. A t(2;5)(p23;q35) chromosomal translocation occurs in approximately 8% of non-Hodgkin lymphomas in children and young adults and results in the chimeric fusion of *NPM1* to *ALK* (17). In rare cases of acute promyelocytic leukemia (APL), characterized by chromosomal translocations that disrupt the gene encoding the retinoic acid receptor (*RARa*), fusion of *NPM1* to *RARa* was shown (18). A t(3;5)(q25.1;q34) chromosomal translocation, infrequently seen in myelodysplastic syndrome and AML gives rise to a fusion transcript of *NPM1* and *MLF1* (19).
Gene expression profiling is a powerful way to comprehensively classify individuals with AML and to further resolving the heterogenous nature of AML (20). Using this technique, new prognostically relevant AML subtypes have been identified, while the presence of recurrent chromosomal abnormalities such as inv(16), t(15;17) and t(8;21) as well as other molecular aberrations, e.g., C- and N-terminal mutations in *CEBPa*, could be predicted with high accuracy by unique expression patterns (21-23). In recent study, novel subtypes of AML have also been defined based on gene expression profiling, however, the common molecular abnormalities in these AML subtypes are largely unknown (22). Since *NPM1* is mutated in approximately one third of AML patients, this molecular abnormality may drive the clustering of these AML subtypes. The effect of mutant NPM1 has been studied using gene expression profiling and revealed a distinctive signature of *NPM1* mutations (24). Amongst players in this signature were several homeodomain-containing family members of HOX transcription factors and CD34, both observations being indicative of hematopoietic development (24). However, it is currently not known whether NPM1 mutations are predictable on the basis of a gene expression signature.
Cytoplasmic NPM1 has been positively associated with remission rate (13), however, the relation of mutant NPM1 with survival outcome parameters remains to be elucidated.
We have studied a well-characterized cohort of 275 cases of *de novo* AML for the presence of a *NPM1* mutations to (I) validate dHPLC as a rapid approach to determine *NPM1* mutations, (II) investigate the relation of *NPM1* mutations with regard to clinical parameters, cytogenetics and various molecular abnormalities, (III) determine the relation of *NPM1* mutations in subtypes of AML, recently identified by gene expression profiling (22), (IV) derive *NPM1* mutation specific and predictive gene expression signatures and (V) determine the prognostic value of mutated *NPM1*.

4.3 Patients and methods

4.3.1 Patients and cell samples

Patients had a diagnosis of primary AML, confirmed by cytological examination of blood and bone marrow (median age 44 (range 15-78), median bone marrow blast count 65 percent (range 0 (for APL) -98), median white blood cell (WBC) count 32 ($x10^9$/l) (range 0.3-263)). All patients had been treated according to the HOVON (Dutch-Belgian Hematology-Oncology Co-operative group) protocols (http://www.hovon.nl) (25-27). After informed consent, bone marrow aspirates or peripheral blood samples were taken at diagnosis. Blasts and mononuclear

cells were purified by Ficoll-Hypaque (Nygaard, Oslo, Norway) centrifugation and cryopreserved. The AML samples contained 80-100 percent blast cells after thawing, regardless of the blast count at diagnosis.

4.3.2 PCR, WAVE and sequence analyses

RNA isolation and cDNA synthesis were performed as described (22, 28). cDNA prepared from 50ng of RNA was used for all PCR amplifications. *NPM1* mutations in exon 12 were determined by cDNA amplification using the primers NPM1-FOR 5'-CTTCCGGATGACTGACCAAGAG-3' and primer NPM1-REV 5'-CCTGGACAACATTTATCAAACACG-3' (25mM dNTP, 15 pmol primers, 2mM $MgCl_2$, Taq polymerase and 10xbuffer (Invitrogen Life Technologies, Breda, The Netherlands)). Cycling conditions for *NPM1* mutation detection were as follows: 1 cycle 5' 94^0C, 30 cycles 1' 94^0C, 1' 58^0C, 1' 72^0C, and 1 cycle 7' 72^0C. PCR products were subsequently subjected to dHPLC using a Transgenomics (Omahah, NE) WAVE dHPLC system (29). Samples were run at 56^0C and 58^0C. The exact *NPM1* mutant sequence was confirmed for all samples showing an abnormal dHPLC profile. PCR products were purified using the Multiscreen-PCR 96-well system (Millipore, Bedford, MA) followed by direct sequencing with NPM1-REV using an ABI-PRISM3100 genetic analyzer (Applied Biosytems, Foster City, CA).
Sequence analyses for mutations in *FLT3* (ITD and tyrosine kinase domain mutation (TKD)), N-*RAS*, K-*RAS* and *CEBPa* performed as described previously.(12, 30, 31).

4.3.3 Gene expression profiling and unsupervised cluster analyses

285 AML cases were analyzed using Affymetrix HGU133A GeneChips (22). Unsupervised cluster analysis on the basis of the gene expression profiles of the 285 cases of AML was performed using the Correlation Visualization tool of Omniviz (Maynard, MA (version 3.6)) (22). The Pearson's correlation values calculated in Omniviz were subsequently imported into the MicroArray Data Explorer (MADEx), which was developed in our laboratory. MADEx was used to visualize the relations between the Omniviz unsupervised clustering results and other parameters, such as clinical and molecular characteristics of the AML patients (Figure 1). MADEx is a database system that stores, mines and visualizes microarray data in a secure and scalable manner.

4.3.4 Significance Analysis of Microarrays (SAM)

All supervised analyses were performed using Significance Analysis of Microarrays (SAM) (version 1.21) (32). A threshold was set for a minimum change in expression of at least 1.5-fold. A false discovery rate of 0.01 was used to select the differentially expressed genes.

4.3.5 Prediction Analysis of Microarrays (PAM)

All supervised class prediction analyses were performed by applying Prediction Analysis of Microarrays (PAM) (version 2.0) (33). The gene signature was selected based on the smallest prediction error in the training set and was subsequently tested using the test set. The positive predictive value was calculated with (true positives/(true positives + false positives)).

4.3.6 Statistical analyses of survival

Cytogenetic abnormalities were categorized in 3 cytogenetic groups for statistical analyses. Patients with inv(16)/t(16;16), t(8;21), and t(15;17) abnormalities were considered as being in the favorable-risk category. The unfavorable-risk category was defined by the presence of -5/del(5q), -7/del(7q), t(6;9), t(9;22), 3q26 abnormality or complex karyotype (more than 3 abnormalities). All other patients were classified as intermediate risk. Statistical analyses were performed with Stata Statistical Software, Release 7.0 (Stata, College Station, TX). Actuarial probabilities of overall survival (OS) (with failure defined as death due to any cause) and event-free survival (EFS) (with failure defined as not achieving complete remission (set at day 1), relapse, or death in first complete remission) were estimated by the method of Kaplan and Meier. The Cox proportional hazards model was applied to determine the association of *NPM1* mutation with OS, EFS and disease-free survival (DFS), without and with adjustment for other factors such as cytogenetic risk, age, white blood cell count and *FLT3* ITD. All tests were 2-sided, and a *P* of less than .05 was considered statistically significant.

4.4 Results

4.4.1 Different *NPM1* variant mutations in AML

The presence of *NPM1* mutations in 275 cases of primary AML was rapidly and reliably detected by dHPLC WAVE. Nucleotide sequencing was performed on those cases with an abnormal dHPLC profile (Table 1). Each *NPM1* mutation variant reveals a specific dHPLC WAVE profile. Thus, each type of *NPM1* mutation could be predicted on the basis of a specific dHPLC WAVE profile.
In addition, three novel *NPM1* mutant variants were identified (*NPM1* mutants I to K (Table 1)). These rare variants have comparable 4-base pair insertions, like *NPM1* variant mutations A to D (13), resulting in a frame shift and replacement of the 7 C-terminal amino acids of the NPM1 protein by 11 different residues (Table 1).

4.4.2 *NPM1* mutations in relation to clinical and molecular features in AML

The *NPM1* mutation frequencies of the 275 cases of primary AML with regard to clinical parameters, morphology, cytogenetics and molecular characteristics are shown in Table 2. *NPM1* mutations are significantly less frequently present in patients of younger age (<35 yr, $p<0.001$). The mean age of patients with AML and *NPM1* mutations is 47.3 (±10.7), whereas the mean age of patients with AML and wild-type *NPM1* is 39,7 (±13.3). *NPM1* mutations are seen in AML FAB subtypes M1 to M6, but are absent in AML FAB M0. The mutations are relatively frequently found in AML FAB M5 as well as in all three cases of AML FAB M6. In AML with recurrent translocations, i.e., t(8;21), inv(16) and t(15;17), no *NPM1* mutations were demonstrated. In cases with various other cytogenetic abnormalities, mutations in *NPM1* were also rare. As a result, there appears a positive correlation between *NPM1* mutations and AML with normal karyotypes ($p<0.001$). The analysis of *NPM1* mutations reveals interesting relationships with particular common molecular abnormalities. The presence of *NPM1* mutations significantly correlates with the presence of *FLT3* ITD mutations ($p<0.001$). A correlation between *NPM1* mutations and *FLT3* TKD mutations is not apparent. AML with N-*RAS* mutations

NPM1 variant	Nucleotide sequence*	Protein sequence†	No. of *NPM1* mutants (%)
WT	GATCTCTG GCAGTGGAGGAAGTCTCTTTAAGAAAATAG	-DLWQWRKSL	NA
Mutant A	GATCTCTG*TCTG*GCAGTGGAGGAAGTCTCTTTAAGAAAATAG	-DLCLAVEEVSLRK	72(26)
Mutant B	GATCTCTG*CATG*GCAGTGGAGGAAGTCTCTTTAAGAAAATAG	-DLCMAVEEVSLRK	12(4)
Mutant D	GATCTCTG*CCTG*GCAGTGGAGGAAGTCTCTTTAAGAAAATAG	-DLCLAVEEVSLRK	4(1)
Mutant I‡	GATCTCTG*CAGA*GCAGTGGAGGAAGTCTCTTTAAGAAAATAG	-DLCRAVEEVSLRK	1(<1)
Mutant J‡	GATCTCTG*CTTG*GCAGTGGAGGAAGTCTCTTTAAGAAAATAG	-DLCLAVEEVSLRK	1(<1)
Mutant K‡	GATCTCTG*TATG*GCAGTGGAGGAAGTCTCTTTAAGAAAATAG	-DLCMAVEEVSLRK	1(<1)

*Inserted nucleotides are underlined and italicized.
†The C-terminal tryptophane residues (W) are underlined for wild-type (WT) NPM1.
‡NPM1 mutants I, J, and K are novel variants identified in addition to the 6 known NPM1 variants (13).

Table 1. NPM1 mutation frequencies in 275 cases of de novo AML. Number of NPM1 mutants with undetermined variants, 4 (1%); total number of NPM1 mutants, 95 (35%).

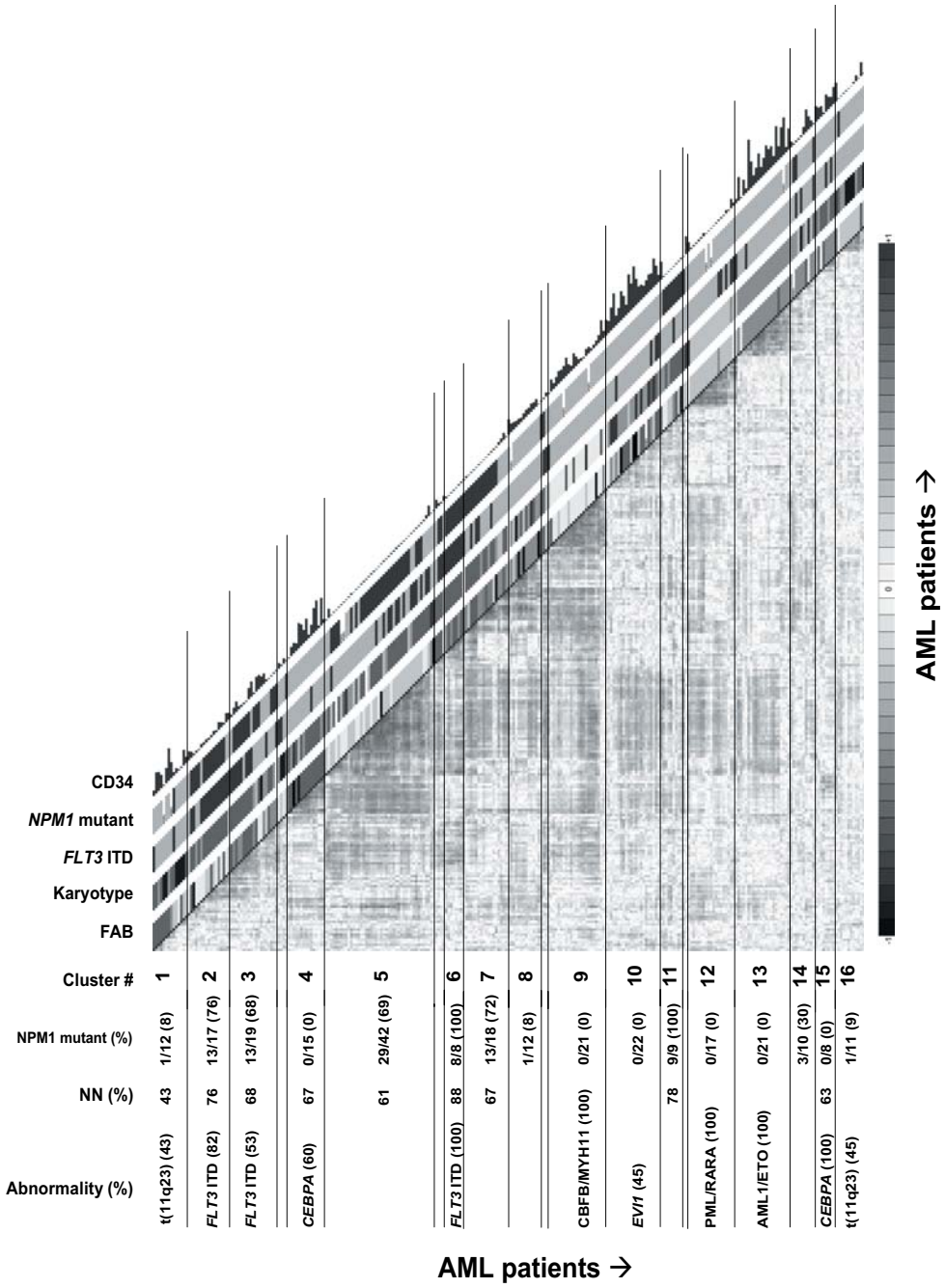
	No.	No. of *NPM1* mutants (%)	*P*
Sex			.100
Male	135	40 (30)	
Female	140	55 (39)	
Age, y			< .001
Younger than 35	74	11 (15)	
35 to 60	169	69 (14)	
60 and older	32	15 (47)	
WBC count, x 10^9/L			<.001
20 or below	113	28 (25)	
Above 20	157	66 (42)	
ND	5	1 (20)	
FAB			
M0	6	0 (0)	ND
M1	62	21 (34)	> .999
M2	63	18 (29)	.296
M3	17	1 (6)	.008
M4	49	15 (31)	.062
M5	65	32 (49)	.007
M6	3	3 (100)	ND
Cytogenetic abnormalities*			
t(15;17)	16	0 (0)	.002
t(8;21)	21	0 (0)	< .001
inv(16/t(16;16)	17	0 (0)	.001
+8	24	5 (21)	ND
+11	5	0 (0)	ND
+21	1	1 (100)	ND
-5	2	0 (0)	ND
-5(q)	1	0 (0)	ND
-7	13	0 (0)	.005
-7(q)	7	0 (0)	ND
3q	5	1 (20)	ND
t(6;9)	4	0 (0)	ND
t(9;22)	2	1 (50)	ND
t(11q23)	16	1 (6)	.014

	No.	No. of *NPM1* mutants (%)	*P*
Complex; more than 3 abnormalities	11	0(0)	.018
Other	55	10 (18)	.004
Normal	116	74 (64)	< 0.001
ND	10	6 (60)	ND
Molecular abnormalities			
FLT3 ITD	78	47 (60)	< .001
FLT3 TKD	32	14 (44)	.243
NRAS	25	3 (12)	.024
KRAS	8	5 (63)	.130
CEBPa	17	0 (0)	.001

P values were calculated using the Fisher exact test (2-tailed; ND indicates not determined).
**All patients with a specific abnormality were considered irrespective of the presence of additional abnormalities.*

Table 2. NPM1 mutation frequencies in relation to clinical parameters, morphology, cytogenetics, and molecular characteristics of the 275 patients with de novo AML.

Figure 1 (facing page). Adapted Omniviz Correlation View of 285 AML patients (2856 probe sets) (22). *The Correlation View displays pair-wise correlations between AML patients. The cells in the visualization are colored by Pearson's correlation coefficient values with deeper colors indicating higher positive (red) or negative (blue) correlations. The scale bar indicates maximum positive correlation (red) towards maximum negative correlation (blue). The 16 clusters identified in the cohort of 285 AML patients on the basis of the Correlation View are indicated (1 to 16) (22). Clinical and molecular data are depicted in the columns along the original diagonal of the Correlation View (22). FAB classification and karyotype based on cytogenetics are depicted in the first two columns (FAB M0-red, M1-green, M2-purple, M3-orange, M4-yellow, M5-blue, M6-grey; karyotype: normal-green, inv(16)-yellow, t(8;21)-purple, t(15;17)-orange, 11q23 abnormalities-blue, 7(q) abnormalities-red, +8-pink, complex-black, other-gray). FLT3 ITD and NPM1 mutations are depicted in the same set of columns (red bar: positive and green bar: negative). The expression levels of CD34 (probe set: 209543_s_at) in the 285 AML patients are plotted in the last column (bars are proportional to the level of expression). The percentages of the most common (>40 percent) abnormalities, NPM1 mutations as well as normal karyotypes (NN) for each cluster are indicated.*
A full-color version of this figure is provided on the CD.


CD34
NPM1 mutant
FLT3 ITD
Karyotype
FAB
Cluster #
NPM1 mutant (%)
NN (%)
Abnormality (%)
t(11q23) (43) 43 1/12 (8) 1
FLT3 ITD (82) 76 13/17 (76) 2
FLT3 ITD (53) 68 13/19 (68) 3
CEBPA (60) 67 0/15 (0) 4
61 29/42 (69) 5
FLT3 ITD (100) 88 8/8 (100) 6
67 13/18 (72) 7
1/12 (8) 8
CBFB/MYH11 (100) 0/21 (0) 9
EVI1 (45) 0/22 (0) 10
78 9/9 (100) 11
PML/RARA (100) 0/17 (0) 12
AML1/ETO (100) 0/21 (0) 13
3/10 (30) 14
CEBPA (100) 63 0/8 (0) 15
t(11q23) (45) 1/11 (9) 16
AML patients →
AML patients →

Probe set id	Gene symbol	Fold change
Up-regulated in AML with mutant *NPM1*		
213844_at	*HOXA5*	4.2
205366_s_at	*HOXB6*	2.6
208414_s_at	*HOXB3*	1.8
204082_at	*PBX3*	2.8
205600_x_at	*HOXB5*	2.1
206289_at	*HOXA4*	2.1
205453_at	*HOXB2*	2.1
213150_at	*HOXA10*	2.6
204069_at	*MEIS1*	2.6
209905_at	*HOXA9*	2.8
201664_at	*SMCL4*	2.1
20943_s_at	*PHKA2*	1.6
206847_s_at	*HOXA7*	1.5
63825_at	*ABHD2*	1.6
219304_s_at	*PDGFD*	1.9
212820_at	*RC3*	2.4
213110_s_at	*COL4A5*	2.9
207111_at	*EMR1*	2.1
208557_at	*HOXA6*	1.6
203471_s_at	*PLEK*	1.6
203680_at	*PRKAR2B*	2.3
202729_s_at	*LTBP1*	3.6
210145_at	*PLA2G4A*	1.6
220162_s_at	*CARD9*	1.5
206298_at	*ARHGAP22*	2.0
219602_s_at	*FAM38B*	1.9
Down-regulated in AML with mutant *NPM1*		
209543_s_at	*CD34*	-5.4
206896_s_at	*GNG7*	-1.7
209583_s_at	*MOX2*	-3.1
200953_s_at	*CCND2*	-2.1
221004_s_at	*ITM2C*	-3.0
205330_at	*MN1*	-4.3
200602_at	*APP*	-2.6

Probe set id	Gene symbol	Fold change
200665_s_at	*SPARC*	-3.8
201015_s_at	*JUP*	-3.1
218899_s_at	*BAALC*	-4.3
209679_s_at	*LOC57228*	-1.9
219694_at	*FLJ11127*	-2.1
206042_x_at	*SNRPN*	-2.5
211535_s_at	*FGFR1*	-2.6
214582_at	*PDE3B*	-1.8
221523_s_at	*RRAGD*	-1.8
213618_at	*CENTD1*	-1.7
218589_at	*P2RY5*	-3.3
202016_at	*MEST*	-2.9
208116_s_at	*MAN1A1*	-3.1
205240_at	*GPSM2*	-1.9
202747_s_at	*ITM2A*	-4.1
206622_at	*TRH*	-9.1
206726_at	*PGDS*	-4.7

Table 3. NPM1 mutation-associated gene expression in 275 patients with de novo AML*. The top 50 unique most discriminating genes and fold change in expression with regard to NPM1 mutation as determined by SAM (up-regulated: increased expression in NPM1 mutant cases; down-regulated: decreased expression in NPM1 mutant cases).*

4.4.3 *NPM1* mutations occur within specific AML subtypes defined by gene expression profiling

Of the cohort of 285 cases of primary AML which had previously been profiled using the Affymetrix HGU133 GeneChip (22) and for which 16 distinct expression signatures had been defined following unsupervised cluster analyses, we have now examined 275 cases for the presence of *NPM1* mutations. Among these pre-established signatures, the AML cases with *NPM1* mutations aggregate within particular clusters (Figure 1). The majority of AML cases in clusters #2, #3, #5 and #7 carry *NPM1* mutations. As a matter of fact, all cases of AML of clusters #6 (100% *FLT3* ITD) and #11 (78% normal karyotypes) carried mutations in *NPM1*. Although, clusters #7 and #8 have comparable expression profiles (Figure 1), 13 out of the 18 AML cases in cluster #7 (72%) and only 1 of 12 cases of cluster #8 (8%) reveal *NPM1* mutations. The clusters merely consisting of AML with inv(16) (#9), t(15;17) (#12) or t(8;21) (#13) as well as the clusters predominantly containing cases with *CEBPa* mutations (#4 and #15) all lack *NPM1* mutations. The subset of AML patients in cluster #10, with adverse prognosis and an expression profile comparable to CD34-positive cells (22), did not present with *NPM1* mutations.

Falini and colleagues (13) had shown a negative correlation between the presence

	Predicted genotype			
	10-CV error*		Error validation set*	
Genotype	**WT**	**Mutant**	**WT**	**Mutant**
WT *NPM1*	98	24	48	10
Mutant *NPM1*	0	62	0	33

**10-fold CV error: 10-fold cross validation prediction error on training set (n=184), Error validation set: prediction error on validation set (n=91). 10-Fold cross validation works as follows: the model is fitted on 90% of the samples and the class of the remaining 10% is predicted. This procedure is repeated 10 times, with each part playing the role of the test samples and the error of all 10 parts added together to compute the overall error. The error within the validation set reflects the number of samples wrongfully predicted in this set.*

Table 4. NPM1 mutation prediction by using PAM. *Most optimal result for NPM1 mutation prediction using a cohort of 275 cases of AML divided in a training and test set (22). 22 probe sets were used in this prediction, representing 18 unique genes.*

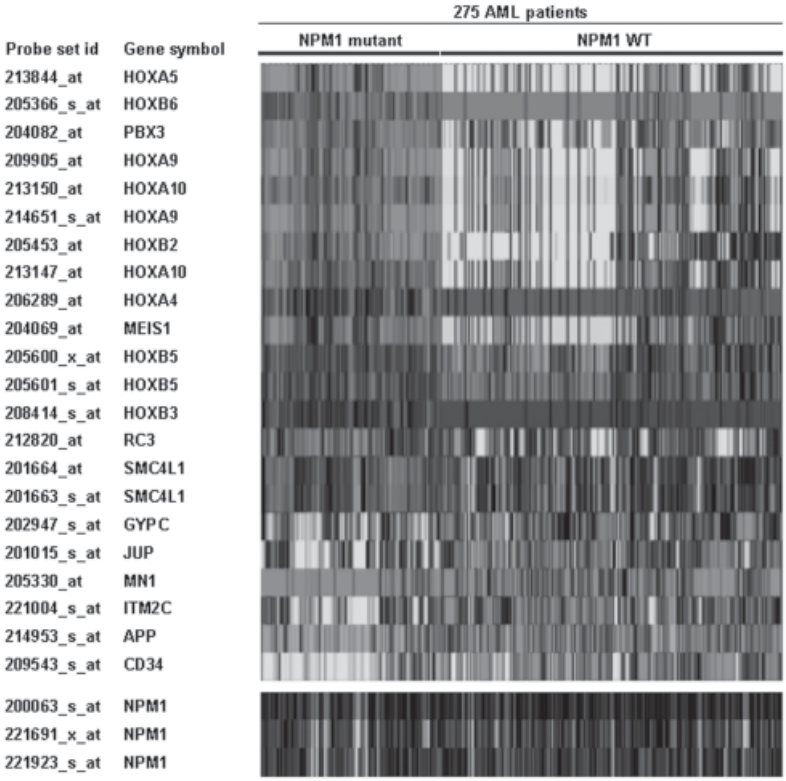


Figure 2. The most predictive molecular signature with regard to NPM1 mutation in AML assessed with 22 probe sets representing 18 unique genes. *The levels of expression of the probe sets in the 275 cases of AML are depicted (scale bar indicates an increase (red) or a decrease (green) in the level of expression of at least 4 relative to the geometric mean) (22). The three probe sets representing the NPM1 gene are also depicted.* ***A full-color version of this figure is provided on the CD.***

of *NPM1* mutations and *CD34* expression levels. In fact, by plotting the *CD34* mRNA expression levels of the 285 AML cases, as determined by the GeneChip analyses, along the *NPM1* mutation status (Figure 1), a distinguishable association of *CD34* mRNA expression and *NPM1* mutation is apparent, i.e., *CD34* mRNA expression is low or absent in cases of AML with *NPM1* mutations, while *CD34* mRNA levels are high in AML cases without *NPM1* mutations.

4.4.4 *HOX* gene-specific gene expression signature of *NPM1* mutant cases

To identify genes with significant differential expression between primary AML samples with *NPM1* mutations (n=95) and samples without *NPM1* mutations (n=180) the SAM algorithm was used (32). A fold change threshold of 1.5 for upregulation of gene expression and 0.667 for downregulation of gene expression was applied. With a False Discovery Rate (FDR) of 0.01, 569 probe sets representing 440 unique genes, were identified as being significantly differentially expressed (Supplementary Figure 1 and Table 1). The identity of the Top-50 genes, in some cases represented by multiple probe sets, are depicted in Table 3.
A dominant homeobox (*HOX*) gene-specific signature is strongly associated with AML carrying an *NPM1* mutation. Moreover, the expression of members of the *HOXA*- and *HOXB*-gene families, but also the *HOX* gene-related TALE genes, *PBX3* and *MEIS1* are increased. In contrast, the *CD34* gene is the strongest significantly down regulated gene with regard to *NPM1* mutation in the AML patient cohort.

4.4.5 *NPM1* mutant cases are predicted with high accuracy based on their gene expression signature

*NPM*1 mutation prediction analyses were performed using the PAM algorithm (33). All 275 primary AML samples were randomly assigned to a training set, consisting of 122 samples without *NPM1* mutations and 62 samples with *NPM1* mutations, and a validation series, consisting of 58 samples lacking the *NPM1* mutation and 33 samples with mutations in *NPM1*. Cross validation to predict the mutation status of *NPM1* on the training set resulted in 100% correct predictions on presence of mutation (sensitivity) and 80.3% correct predictions on absence of mutation (specificity) (Table 4). Prediction of an independent validation set also resulted in 100% correct prediction of presence of mutation and 82.7% correct prediction on absence of mutation. The positive predictive value in this cohort is 72% in the training set and 74% overall. As expected, the genes included in the PAM gene signature were among the most significant differentially expressed genes as determined by SAM, thereby validating both algorithms. Of note, *NPM1* mRNA expression did not correlate with mutation status (Figure 2). The *NPM1* mutant AML cases have a distinct signature with regard to the 18 selected genes (Figure 2) and are therefore predicted with high accuracy (Table 4). However, these 18 genes seem to be expressed at comparable levels in a subset of AML cases with wild type *NPM1* (Figure 2).

4.4.6 *NPM1* mutation is an independent favorable prognostic marker

To investigate the prognostic value of *NPM1* mutation 252 AML patients with long term follow up following therapy completion were included for survival analysis.

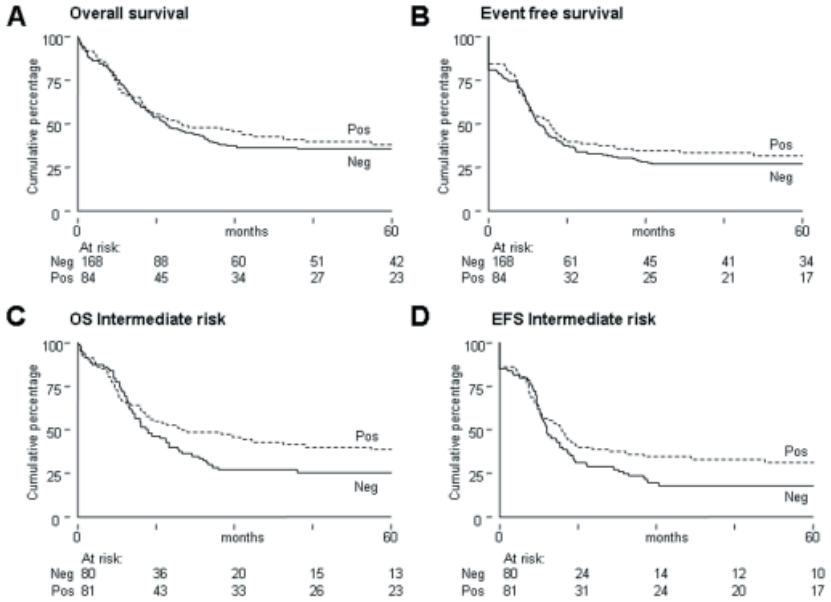


Figure 3. Survival analyses of AML patients with and without NPM1 mutations. *Kaplan-Meier estimates of overall survival (A) and event free survival (B) among patients with AML, overall survival (C) and event free survival (D) among patients with AML with intermediate risk karyotypes.*

The EFS, OS and probability of relapse at 60 months for the AML patients with or without *NPM1* mutations were similar (Figures 3A and 3B). Likewise, among the subgroup with cytogenetics of intermediate prognostic risk EFS, OS and probability of relapse at 60 months were not different either (Figures 3C and 3D), although there appears a trend for more favorable outcome for patients with AML with *NPM1* mutations. Since *NPM1* mutations are significantly associated with both normal karyotype and presence of a *FLT3* ITD (Table 2), we wished to investigate the prognostic value of *NPM1* mutations within the intermediate cytogenetic risk group in relation to *FLT3* ITD status. Patients in the intermediate cytogenetic risk group without *FLT3* ITD mutations but with *NPM1* mutations have a significantly better OS and EFS than those without *NPM1* mutations (p=0.05) (Figures 4A and C). In intermediate cytogenetic risk AML with *FLT3* ITD mutations *NPM1* mutations do not significantly distinguish prognosis (Figures 4B and D).

NPM1 mutations are asynchronously associated with particular karyotypes and molecular abnormalities that might express additional positive or negative prognostic value and therefore might obscure the prognostic significance of *NPM1* mutations (Table 2 and Figure 1). Therefore we investigated the prognostic value of *NPM1* mutations in both univariable and multivariable analyses. Univariable and multivariable Cox regression analysis were applied to assess the prognostic

significance of *NPM1* mutation, cytogenetic risk class, WBC below or above 20, age and *FLT3* ITD mutation for OS, DFS and EFS (Table 5). In univariable analyses *NPM1* mutation showed no statistical significance with respect to the endpoints. However, in multivariable analyses mutated *NPM1* was significantly associated with a much lower hazard ratio, which was statistically significant for all endpoints (EFS: HR = 0.59, p = 0.005; DFS: HR = 0.52, p = 0.003 and OS: HR = 0.49, p = 0.0003). Mutant *NPM1* appeared as independent prognostic marker in addition to cytogenetic risk, age and *FLT3* ITD mutations.

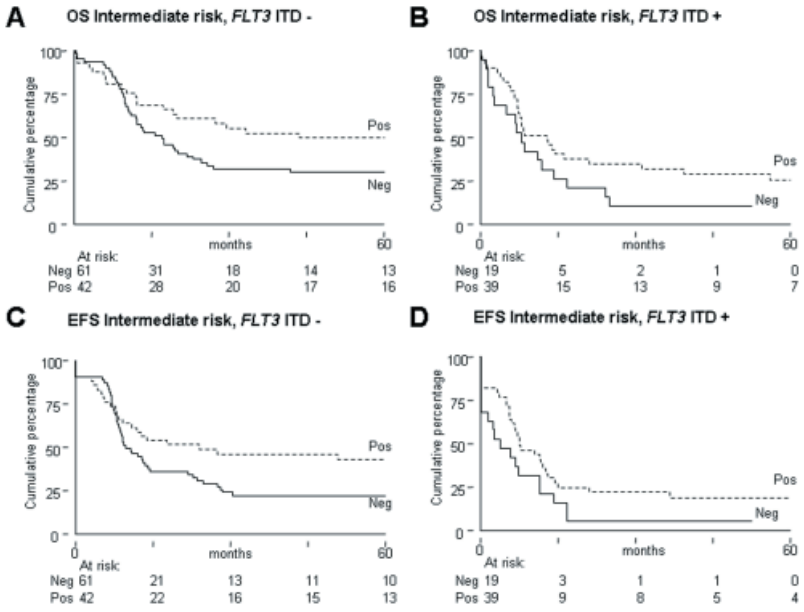


Figure 4. Survival analyses of intermediate cytogenetic risk AML patients with and without FLT3 ITD and/or NPM1 mutations. *Kaplan-Meier estimates of overall survival (A and B) and event free survival (C and D) for patients with intermediate risk AML and FLT3 ITD mutations (FLT3 ITD+ (B and D)) versus those with intermediate risk AML without FLT3 ITD mutations (FLT3 ITD- (A and C))*

	EFS		DFS		OS	
	HR (95% CI)	P*	HR (95% CI)	P*	HR (95% CI)	P*
Univariable						
Intermediate†	2.06 (1.33 -3.20)	.001*	2.82 (1.63-4.87)	< .001*	2.67 (1.62-4.41)	< .001*
Poor†	3.88 (2.29-6.56)	< 0.001*	5.09 (2.61-9.91)	< 0.001*	4.44 (2.49-7.93)	< 0.001*
WBC count‡	1.14 (0.84-1.53)	.39	1.18 (0.84-1.68)	.34	1.19 (0.87-1.62)	.29
Age, decades	1.05 (0.94-1.19)	.38	1.06 (0.92-1.22)	.41	1.17 (1.03-1.32)	.14*
FLT3 ITD§	1.76 (1.28-2.41)	.001*	1.71 (1.17-2.50)	.006*	1.89 (1.36-2.62)	< 0.001*
NPM1 mutation |	0.88 (0.64-1.21)	.43	0.93 (0.64-1.34)	.69	0.90 (0.65-1.27)	.57
Multivariable						
Intermediate†	2.18 (1.36-3.50)	0.001*	3.11 (1.74-5.55)	< 0.001*	2.87 (1.69-4.86)	< 0.001*
Poor†	3.96 (2.34-6.74)	< 0.001*	5.73 (2.93-11.22)	< 0.001*	4.68 (2.60-8.43)	< 0.001*
WBC count‡	1.06 (0.78-1.44)	.72	1.10 (0.77-1.58)	.60	1.18 (0.85-1.63)	.32
Age, decades	1.07 (0.95-1.21)	.27	1.10 (0.95-1.27)	.21	1.19 (1.05-1.35)	.008*
FLT3 ITD	1.96 (1.39-2.76)	< 0.001*	2.00 (1.31-3.06)	0.001*	2.13 (1.49-3.05)	< 0.001*
NPM1 mutation	0.59 (0.41-0.85)	.005*	0.52 (0.34-0.80)	0.003*	0.49 (0.33-0.72)	< 0.001*

CI indicates confidence interval.
**P values .05.*
†Cytogenetic risk versus cytogenetic good risk.
‡More than 20 109/L versus less than 20 109/L.
§FLT3 ITD versus no FLT3 ITD.
|| NPM1 mutation versus no NPM1 mutation.

Table 5. Univariable and multivariable analyses of NPM1 mutation as prognostic factor for EFS, DFS, and OS in AML.

4.5 Discussion

Recently, we established a comprehensive classification of AML based on whole-genome expression profiling (22). In this classification several clusters of AML signatures correlated with distinct (cyto)genetic abnormalities, such as t(8;21), t(15;17), inv(16) and C- and N-terminal mutations in CEBPα. However, the common underlying molecular abnormality for the other subtypes of AML was unknown. In the present study we show that two clusters consist entirely of AML cases with *NPM1* mutations, i.e., clusters #6 and #11, whereas clusters #2, #3, #5 and #7 predominantly include patients with *NPM1* mutations. Thus, we identify mutant *NPM1* as a common molecular abnormality in these subtypes of AML.

Falini and colleagues (13) have shown that mutations in *NPM1* are present in 35% of patients with AML. In this study, we confirm that *NPM1* mutations are frequently present in AML, i.e., in 35% of all cases. *NPM1* mutant is less frequently represented in patients with AML of age less than 35 yr. This seems consistent with the tendency of the *NPM1* mutation to be more frequently present in older children with AML (34). Furthermore, *NPM1* mutations are significantly associated with AML with high WBC.

We detected the various mutation variants of *NPM1* at similar frequencies as was described recently (13), and also identified three novel mutant variants. These novel variants carry, like the other *NPM1* mutant variants in our study, an insertion of 4 bp, resulting in a protein with an altered C-terminus (Table 1). Falini and colleagues (13) suggested that the disruption of one of the two C-terminal tryptophan residues and the last five residues, i.e., VSLRK, are important for *NPM1* mutant function. Our study suggests that the final 9 aminoacids, i.e., AVEEVSLRK, may in general be required for mutant NPM1 function.

NPM1 mutations often coincide with mutations in *FLT3*, in particular with the ITD-type mutations. Our data may suggest an association of mutations in K-*RAS* and *NPM1*, but the study has limited power because of the small number of mutant K-*RAS* AML. In contrast, mutations in N-*RAS* were not associated with mutant *NPM1*, since N-*RAS* mutations are found in AML with inv(16) (22, 30), a subclass of AML which lacks *NPM1* mutations. In addition, we did not find *NPM1* mutations in clusters of AML patients, previously identified by gene expression profiling (22), characterized by C- and N-terminal mutations in CEBPα. These observations might perhaps suggest that constitutive active FLT3 or K-RAS provide the proliferative signal, whereas mutant NPM1 might serve to impair differentiation in the multistep pathogenesis model of AML (35).

It is of note that, the discriminative genes identified by SAM and PAM revealed a strong *HOX*- and *TALE*- gene-specific signature associated with AML cases with mutant *NPM1*, as was published very recently (24). Thus, although CD34 has generally been used as marker for immature hematopoietic progenitor cells, the *NPM1* mutant CD34-negative cells display a molecular signature similar to that of hematopoietic stem cells (HSC). Recent studies have shown that CD34-negative HSCs exist as well (36, 37). These CD34-negative cells also posses HSC-specific characteristics, including the ability for hematopoietic engraftment (36, 37). This brings up the question whether the *NPM1* mutant AML cells in fact represent a more primitive population of HSCs with a *HOX*-gene signature.

The sharp distinction between CD34-positivity and *NPM1* mutations in AML

cases based on the *HOX*-gene signature is notable since 22 of the 39 *HOX* genes are expressed in human CD34-positive cells, whereas during normal differentiation *HOX* gene expression declines (38). Extensive studies have demonstrated for a number of *HOX*-genes that sustained overexpression in murine bone marrow results in perturbations in the stem cell pools, and co-expression of certain *HOX*-gene family members with their protein binding partner, such as MEIS1, results in leukemia (38). Thus, in AML with *NPM1* mutations the hematopoietic progenitor cells may have arrested at a differentiation stage with endogenous co-expression of the *HOX*-genes, i.e., *HOXA5*, *-A9*, *-A10, -B2, -B3, -B5* and *-B6*, and their *TALE* partner genes, i.e., *MEIS1* and *PBX3*, or as a result of the increased expression of these genes.

SAM and PAM analyses were highly concordant for the genes identified with differential expression between AML with mutant and AML with wild type *NPM1*, where, in both cases, *CD34* was the most discriminating gene, downregulated in *NPM1* mutants. Patients harboring an *NPM1* mutation can be predicted with high accuracy, however, a subset of patients with wild type *NPM1* is wrongly predicted. Interestingly, within this subgroup the percentage of patients with 11q23 abnormalities is significantly increased. In fact, this may not be surprising since MLL has been implicated as a *HOX* gene regulator (39) and selective expression of *HOX* genes in ALL cases with mutant MLL has been shown (40).

By univariate analyses we show that there is a tendency that the presence of an *NPM1* mutation, and concomitant low *CD34* mRNA expression, is a favorable marker for clinical outcome (Figure 3). This is in agreement with CD34 expression as an indicator for poor response to induction therapy (41-43). In addition, we demonstrate by multivariate analyses that *NPM1* mutations are a strong independent favorable predictive marker for EFS, DFS and OS in AML. The finding that the effect of *NPM1* mutation is much more pronounced in multivariable than in univariable analyses can be explained by the strong correlation between *NPM1* and *FLT3* ITD mutations, and additionally by the association with high WBC and age. *NPM1* mutations are more frequently present in AML patients with *FLT3* ITD, high WBC and older age. These factors are unfavorable determinants of prognosis (44), while *NPM1* mutation seems to express a favorable prognostic impact. This implies that in univariable analysis the positive effect of *NPM1* mutations, as measured by the method of Kaplan-Meier or by the HR, is less pronounced, since the adverse effects of *FLT3* ITD, high WBC and age might mask this effect. In fact, *NPM1* mutations distinguish a favorable subgroup among intermediate cytogenetic risk *FLT3* ITD negative AML, that shows comparatively better OS and EFS (Figure 4). In addition, we note that in the multivariable analyses the effect of *NPM1* mutations is more pronounced than is apparent in the univariable analyses. This is not only true for *NPM1* mutations, but also for the independent prognostic effect of *FLT3* ITD mutations, age and cytogenetic risk (Table 5).

The data presented here demonstrate that the frequent C-terminal insertion mutations in NPM1 correlate with favorable outcome for patients with AML. Since NPM1 mutations are predominantly found in patients with standard risk AML, determination of the NPM1 mutation status, in combination with other prognostically relevant markers, will be useful for further risk stratification of adult patients with *de novo* AML.

Acknowledgements

We are indebted to our colleagues from the bone marrow transplantation group and molecular diagnostics laboratory for storage of the samples and molecular analyses, respectively.

References

1. Lowenberg B, Downing JR, Burnett A. Acute myeloid leukemia. N Engl J Med 1999, 341(14):1051-1062.
2. Slovak ML, Kopecky KJ, Cassileth PA, et al. Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group Study. Blood 2000, 96(13):4075-4083.
3. Byrd JC, Mrozek K, Dodge RK, et al. Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). Blood 2002, 100(13):4325-4336.
4. Grimwade D, Walker H, Oliver F, et al. The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. Blood 1998, 92(7):2322-2333.
5. Grimwade D, Walker H, Harrison G, et al. The predictive value of hierarchical cytogenetic classification in older adults with acute myeloid leukemia (AML): analysis of 1065 patients entered into the United Kingdom Medical Research Council AML11 trial. Blood 2001, 98(5):1312-1320.
6. Gilliland DG, Griffin JD. The roles of FLT3 in hematopoiesis and leukemia. Blood 2002, 100(5):1532-1542.
7. Levis M, Small D. FLT3: ITDoes matter in leukemia. Leukemia 2003, 17(9):1738-1752.
8. Shiah HS, Kuo YY, Tang JL, et al. Clinical and biological implications of partial tandem duplication of the MLL gene in acute myeloid leukemia without chromosomal abnormalities at 11q23. Leukemia 2002, 16(2):196-202.
9. Dohner K, Tobis K, Ulrich R, et al. Prognostic significance of partial tandem duplications of the MLL gene in adult patients 16 to 60 years old with acute myeloid leukemia and normal cytogenetics: a study of the Acute Myeloid Leukemia Study Group Ulm. J Clin Oncol 2002, 20(15):3254-3261.
10. Barjesteh van Waalwijk van Doorn-Khosrovani S, Erpelinck C, van Putten WL, et al. High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. Blood 2003, 101(3):837-845.
11. Preudhomme C, Sagot C, Boissel N, et al. Favorable prognostic significance of CEBPα mutations in patients with de novo acute myeloid leukemia: a study from the Acute Leukemia French Association (ALFA). Blood 2002, 100(8):2717-2723.
12. van Waalwijk van Doorn-Khosrovani SB, Erpelinck C, Meijer J, et al. Biallelic mutations in the CEBPα gene and low CEBPα expression levels as prognostic markers in intermediate-risk AML. Hematol J 2003, 4(1):31-40.
13. Falini B, Mecucci C, Tiacci E, et al. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. N Engl J Med 2005, 352(3):254-266.
14. Dumbar TS, Gentry GA, Olson MO. Interaction of nucleolar phosphoprotein B23 with nucleic acids. Biochemistry 1989, 28(24):9495-9501.
15. Cordell JL, Pulford KA, Bigerna B, et al. Detection of normal and chimeric nucleophosmin in human cells. Blood 1999, 93(2):632-642.
16. Borer RA, Lehner CF, Eppenberger HM, Nigg EA. Major nucleolar proteins shuttle between nucleus and cytoplasm. Cell 1989, 56(3):379-390.
17. Morris SW, Kirstein MN, Valentine MB, et al. Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma. Science 1994, 263(5151):1281-1284.
18. Redner RL, Rush EA, Faas S, Rudert WA, Corey SJ. The t(5;17) variant of acute promyelocytic leukemia expresses a nucleophosmin-retinoic acid receptor fusion. Blood 1996, 87(3):882-

886.
19. Yoneda-Kato N, Look AT, Kirstein MN, et al. The t(3;5)(q25.1;q34) of myelodysplastic syndrome and acute myeloid leukemia produces a novel fusion gene, NPM-MLF1. Oncogene 1996, 12(2):265-275.
20. Valk PJ, Delwel R, Lowenberg B. Gene expression profiling in acute myeloid leukemia. Curr Opin Hematol 2005, 12(1):76-81.
21. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 2004, 350(16):1605-1616.
22. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 2004, 350(16):1617-1628.
23. Ross ME, Mahfouz R, Onciu M, et al. Gene Expression Profiling of Pediatric Acute Myelogenous Leukemia. Blood 2004.
24. Alcalay M, Tiacci E, Bergomas R, et al: Acute myeloid leukemia bearing cytoplasmic nucleophosmin (NPMc+AML) shows a distinct gene expression profile characterized by up-regulation of genes involved in stem cell maintenance. Blood 2005.
25. Lowenberg B, Boogaerts MA, Daenen SM, et al. Value of different modalities of granulocyte-macrophage colony-stimulating factor applied during or after induction therapy of acute myeloid leukemia. J Clin Oncol 1997, 15(12):3496-3506.
26. Lowenberg B, van Putten W, Theobald M, et al. Effect of priming with granulocyte colony-stimulating factor on the outcome of chemotherapy for acute myeloid leukemia. N Engl J Med 2003, 349(8):743-752.
27. Ossenkoppele GJ, Graveland WJ, Sonneveld P, et al. The value of fludarabine in addition to ARA-C and G-CSF in the treatment of patients with high-risk myelodysplastic syndromes and AML in elderly patients. Blood 2004, 103(8):2908-2913.
28. Van der Reijden BA, de Wit L, van der Poel S, et al. Identification of a novel CBFB-MYH11 transcript: implications for RT-PCR diagnosis. Hematol J 2001, 2(3):206-209.
29. Choy YS, Dabora SL, Hall F, et al. Superiority of denaturing high performance liquid chromatography over single-stranded conformation and conformation-sensitive gel electrophoresis for mutation detection in TSC2. Ann Hum Genet 1999, 63 (Pt 5):383-391.
30. Valk PJM, Bowen DT, Frew ME, Goodeve AC, Löwenberg B, Reilly JT. Second hit mutations in the RTK/RAS signalling pathway in acute myeloid leukaemia and inv(16). Haematologica 2004, 89(01):106.
31. Care RS, Valk PJ, Goodeve AC, Abu-Duhier FM, et al. Incidence and prognosis of c-KIT and FLT3 mutations in core binding factor (CBF) acute myeloid leukaemias. Br J Haematol 2003, 121(5):775-777.
32. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001, 98(9):5116-5121.
33. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002, 99(10):6567-6572.
34. Cazzaniga G, Dell'oro MG, Mecucci C, et al. Nucleophosmin mutations in childhood acute myelogenous leukemia with normal karyotype. Blood 2005.
35. Gilliland DG: Molecular genetics of human leukemias. new insights into therapy. Semin Hematol 2002, 39(4 Suppl 3):6-11.
36. Engelhardt M, Lubbert M, Guo Y. CD34(+) or CD34(-): which is the more primitive? Leukemia 2002, 16(9):1603-1608.
37. Guo Y, Lubbert M, Engelhardt M. CD34- hematopoietic stem cells: current concepts and controversies. Stem Cells 2003, 21(1):15-20.
38. Grier DG, Thompson A, Kwasniewska A, McGonigle GJ, Halliday HL, Lappin TR. The pathophysiology of HOX genes and their role in cancer. J Pathol 2005, 205(2):154-171.
39. Daser A, Rabbitts TH. The versatile mixed lineage leukaemia gene MLL and its many associations in leukaemogenesis. Semin Cancer Biol 2005, 15(3):175-188.
40. Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002, 30(1):41-47.
41. Chang H, Salma F, Yi QL, Patterson B, Brien B, Minden MD. Prognostic relevance of immunophenotyping in 379 patients with acute myeloid leukemia. Leuk Res 2004, 28(1):43-48.
42. Myint H, Lucie NP. The prognostic significance of the CD34 antigen in acute myeloid

leukaemia. Leuk Lymphoma 1992, 7(5-6):425-429.
43. Raspadori D, Lauria F, Ventura MA, et al. Incidence and prognostic relevance of CD34 expression in acute myeloblastic leukemia: analysis of 141 cases. Leuk Res 1997, 21(7):603-607.
44. Lowenberg B. Prognostic factors in acute myeloid leukaemia. Best Pract Res Clin Haematol 2001, 14(1):65-75.

Chapter 5

Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia

Roel G. W. Verhaak, Stefan J. Erkeland, Peter J.M. Valk,
Ruud Delwel, Bob Löwenberg and Ivo P. Touw

Department of Hematology, Erasmus Medical Center, Rotterdam,
The Netherlands

Cancer Research. 2006 Jan 15;66(2):622-6.

5.1 Abstract

Retroviral insertion mutagenesis is considered a powerful tool to identify cancer genes in mice, but its significance for human cancer has remained elusive. Moreover, it has recently been debated whether common virus integrations are always a hallmark of tumor cells and contribute to the oncogenic process. Acute myeloid leukemia (AML) is a heterogeneous disease with a variable response to treatment. Recurrent cytogenetic defects and acquired mutations in regulatory genes are associated with AML subtypes and prognosis. Recently, gene expression profiling (GEP) has been applied to further risk-stratify AML. Here, we show that mouse leukemia genes identified by retroviral insertion mutagenesis are more frequently differentially expressed in distinct subclasses of adult and pediatric AML than randomly selected genes or genes located more distantly from a virus integration site. The candidate proto-oncogenes showing discriminative expression in primary AML could be placed in regulatory networks mainly involved in signal transduction and transcriptional control. Our data support the validity of retroviral insertion mutagenesis in mice for human disease and indicate that combining these murine screens for potential proto-oncogenes with GEP in human AML may help to identify critical disease genes and novel pathogenetic networks in leukemia.

5.2 Introduction

Retroviral insertion mutagenesis in mice is used to discover genes involved in leukemia and lymphoma (1). Recent advances in high through-put sequencing and genome-wide BLAST searches and methods to amplify genomic sequences flanking the virus integration site (VIS) resulted in a catalogue of potential cancer genes (2-6). VIS-flanking genes in independent tumors, i.e., common VIS or CIS genes, are considered *bona fide* disease genes. VIS genes not yet found common often also belong to gene classes associated with cancer and may qualify as disease genes (2, 4, 6, 7). Finally, genes located more distantly from a virus integration may also be deregulated and contribute to disease, but the likelihood of this is unknown (7). Some genes identified in murine screens have been implicated in human cancer, but for the majority this has not yet been demonstrated. Moreover, it has recently been debated whether clustering of proviral insertions, previously considered a hallmark of cancer-related integrations, are selected for during the oncogenic process, or to a significant extent reflect the nonrandom nature of integrations in the genome not necessarily linked with tumor outgrowth (7). To establish their significance for clinical disease, we studied expression of VIS and CIS genes in human AML. Gene expression profiling (GEP) has highlighted the heterogeneous nature of human AML and resulted in the identification of leukemia subsets based on gene expression signatures (8-10). Here, we show that VIS genes from different leukemia models contribute significantly to the expression signatures of both adult and pediatric AML. In contrast, no significant correlations were found with the 2 adjacent genes of the VIS or with other genes within a distance of 1Mb, suggesting that genes directly flanking the virus integrations are the principle candidate disease genes. Finally, we provide data suggesting that regulatory networks, predicted by the VIS genes, may discriminate between biologically distinct AML subsets.

5.3 Materials and methods

5.3.1 GEP data from AML patients

Data from Affymetrix HGU133A GeneChip analysis in 285 adult AML patients are available at http://www.ncbi.nlm.nih.gov/geo/, accession number GSE1159 (10). Patients were categorized as favorable and unfavorable risk based on cytogenetic parameters. The favorable risk group comprised cases with t(8;21), inv(16) and t(15;17) without additional unfavorable cytogenetic abnormalities. The unfavorable group comprised cases with complex karyotype abnormalities, -5 or 5q-, -7 or 7q-, t(6;9), t(9;22), 11q23 abnormalities but no favorable risk abnormalities. GEP data from 130 childhood AML samples are available at website http://www.stjuderesearch.org/data/AML1/(9). Data were normalized by global scaling (Affymetrix Microarray Suite version 5.0; MAS5.0) with target average intensity values of 100 for the adult AML and 500 for the pediatric AML dataset, respectively (9, 10). Because these methods reliably identify signals with average intensity values above 30 and 100, minimum thresholds were set at those values for the adult and pediatric AML data sets, respectively. Expression levels from each probe set in every sample were calculated relative to the geometric mean and logarithmically transformed (base 2) to ascribe equal weight to gene-expression levels with equal relative distance to the geometric mean. Significance analysis of microarrays (SAM) (11) was used to identify genes contributing to the unsupervised clustering of patients in the adult and pediatric AML groups. In the adult AML dataset, 16 classes resulting from unsupervised clustering (10) were evaluated. In the juvenile AML dataset, 5 classes defined according to cytogenetic aberrations (9) were analyzed. This analysis was performed for all probe sets represented on the HGU133A GeneChip (n=22283). Patients from a specific class were tested compared to all remaining samples using an S-test and sample-class permutations to assess statistical significance. Probe sets were considered differential when Fold Change values exceeded 1.5 or were lower than 0.67, scores were over 4 or less than -4 and q-values were below 0.05, where False Discovery Rate was lower than 0.05.

5.3.2 Significance of difference in number of differentially expressed probe sets

To calculate the significance of difference in the number of differentially expressed probe sets in two groups, i.e. VIS representing probe sets versus probe sets not representing a VIS, Pearson's Chi-square with one degree of freedom was calculated using 2x2 contingency tables. As some probe sets were differential in multiple clusters, all possibilities on differential expression were taken into account. For instance, 16 SAM analyses were performed on the adult AML dataset; therefore the sum of the numbers used in the contingency table was 16 * 22283 (the total number of probe sets). All occurrences of differential expression were counted, meaning that if a probe set is differential in n clusters, it is counted n times.

5.3.3 Virus flanking genes in mouse leukemia

Genes affected by virus integrations in Graffi 1.4 (Gr-1.4), BXH2 and AKxD murine leukemia virus (MuLV) models have been previously reported (3, 12) (http://genome2.ncifcrf.gov/RTCGD/).

5.3.4 Network and principal component analyses

Ingenuity Pathway Analysis (http://www.ingenuity.com) was used in combination with the Ingenuity Pathways Knowledge Base (IPKB). Genes selected from experimental data, called focus genes, are used for the generation of networks with a maximal size of 35 genes/proteins. Focus genes were VIS genes that significantly contributed to the unsupervised clustering of 285 AML cases. Principal component analysis (PCA) was performed (Spotfire, Inc. Somerville, MA).

5.4 Results

5.4.1 Gr-1.4 VIS genes and adult AML

To assess the relevance of Gr-1.4 VIS and CIS genes for human AML, we determined their expression in different classes of adult AML patients (9, 10). Based on unsupervised cluster analysis of GEP data, 285 adult AML cases were grouped in 16 subclasses (10). With SAM, specific gene sets were linked to these subclasses, by comparing each subclass to the remaining cases. In total, 5193 probe sets, representing 3644 genes, contributed to the signature of the 16 subclasses (Suppl. Table 1a). We calculated that the probability that a randomly selected gene is differentially expressed in one or more subclasses is 0.28 (Table 1) and performed Pearson's Chi-square analysis to test whether VIS and CIS genes have a higher than random probability to be differentially expressed in one of the subclasses. Four gene lists derived from the Gr-1.4-induced leukemia model and represented on the HGU133A GeneChip were analyzed: (I) VIS + CIS genes (n=115, represented by 234 probe sets), (II) CIS genes (n=51, 116 probe sets), (III) direct neighbors of CIS genes (n=53, 81 probe sets) (IV) genes located within a region of 1 Mb of the CIS genes, with a maximum of 5 genes up- or downstream (n=279, 468 probe sets) (Figure 1; Suppl. Table 2a-d). The VIS and CIS genes have a significantly increased probability (0.46, p=000.1 and 0.43, p=0.002, respectively) to be differentially expressed in subclasses of adult AML compared to unselected genes (I and II in Table 1, genes are listed in supplementary Table 3a and 3b). In contrast, no such correlation was found for gene lists III and IV (Table 1).

5.4.2 Gr-1.4 VIS genes and pediatric AML

To determine the validity of these results for an independent AML GEP data set, correlation analysis was performed on 130 childhood AML samples (9). Patients were grouped in 5 subclasses, i.e., cases with inv(16), t(15;17), t(8;21), translocations involving MLL and cases with megakaryoblastic leukemia (Supplementary Table 1b). In total, 2736 probe sets, representing 2093 genes, contributed to the signature of the 5 subclasses. The probability that a randomly selected gene is differentially expressed in one or more subclasses of the childhood AML dataset was 0.16 (Table 1). Similar to adult AML, Gr-1.4 CIS and VIS genes had a significantly increased probability (0.31, p=0.0127 and 0.25, p=0.005, respectively) to be differentially expressed in the distinct patient clusters, while again no such correlation was seen with more distantly located genes (Supplementary Tables 3e and 3f).

	Number of unique genes (probe sets)	Number of unique SAM genes in adult AML (probe sets)	Probability ¥	P-value *	Number of unique SAM genes in pediatric AML (probe sets)	Probability	
All genes	12848 (22283)	3644 (5193[$])	0.28	--	2093 (2736[$])	0.16	--
(I) Gr-1.4 CIS genes	51 (116)	22 (32)	0.43	0.002	16 (21)	0.31	0.0127
(II) Gr-1.4 VIS genes	115 (234)	53 (74)	0.46	<0.0001	29 (42)	0.25	0.0050
(III) 2 adjacent genes	53 (81)	15 (18)	0.28	0.49 (n.s.)	7 (9)	0.13	0.9361 (n.s.)
(IV) 10 adjacent genes	279 (468)	91 (123)	0.33	0.19 (n.s.)	50 (66)	0.18	0.2071 (n.s.)
Candidate leukemia genes from other mouse models							
BXH2 CIS/VIS	53 (111)	33 (51)	0.62	<0.0001	21 (25)	0.40	0.0001
AKxD CIS/VIS	119 (232)	72(104)	0.61	<0.0001	43 (60)	0.36	<0.0001
All CIS/VIS	237 (470)	122 (178)	0.51	<0.0001	69 (97)	0.29	<0.0001

[$] *Because some probe sets contributed to multiple classes, the total number of sets used in Chi-square analysis was 8739 for adult AML and 2955 for the pediatric AML cases. For details see supplementary Tables 1a and 1b.*

** P-value determined by a two-tailed chi-square test with 95% confidence intervals.*

¥ Probability represents the likelihood that a probe set is differentially expressed (number of SAM genes/total number of genes).

n.s: not significant

Table 1. Virus integration sites projected on 285 adult AML and 130 pediatric AML samples

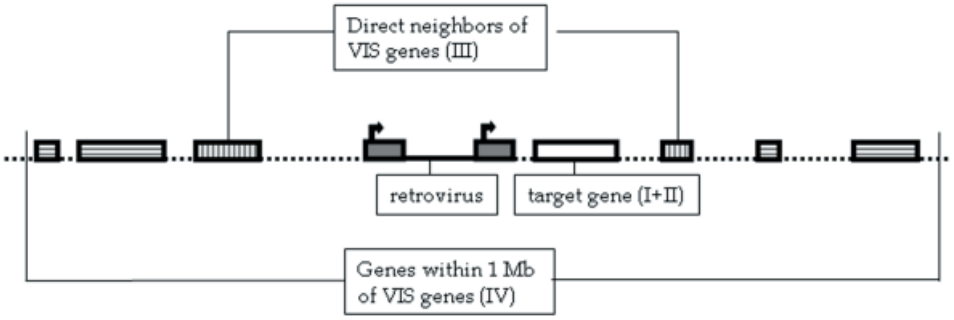


Figure 1. Genomic region of VIS. *Four gene lists were derived from the Gr-1.4–induced leukemia model: (I) genes directly flanking virus integration sites; (II) genes commonly targeted by virus integrations (CIS genes), (III) two direct neighbors of CIS genes, and (IV) genes located within a region of 1 Mbp of CIS genes, with a maximum of 10 (IV). Virus integrations can be located upstream or downstream or within the target gene.*
A full-color version of this figure is provided on the CD.

5.4.3 BXH2 and AKxD virus VIS genes and AML

Candidate leukemia genes identified in two other models, BXH2 and AKxD (Supplementary Tables 2e and 2f; http://genome2.ncifcrf.gov/RTCGD/ also correlated significantly with the gene sets responsible for clustering of adult (0.62, $p<0.0001$ and 0.61, $p<0.0001$, for BXH2 and AKxD CIS/VIS, respectively) and pediatric AML cases (0.40, $p=0.0001$ and 0.36, $p<0.0001$, respectively) (Table 1; Supplementary Tables 3c-f). The combined data from the three models indicate that genes directly flanking the virus integrations are significantly more differentially expressed than random genes in both adult and pediatric AML subtypes.

5.4.4 No correlation between proviral integration and actively transcribed genes in normal hematopoietic precursors

To investigate whether correlations between murine VIS genes and human AML clustering are biased by preferential integrations in genes that are highly expressed in nonleukemic hematopoietic precursors, we calculated the numbers of VIS genes in 5 categories of genes, classified on the basis of their expression levels in normal CD34+ cells (Supplementary Table 4). We found that the greatest portion of integrations occurred in the low to intermediate expression categories and not in highly expressed genes. We also calculated that VIS genes correlated with AML clustering with a significantly higher probability than the nonVIS genes in the different expression categories in CD34+ cells. Together, these results argue against bias due to preferential integration in highly expressed genes (Supplementary Table 5).

	Focus genes in network*	Major global functions of network
Network 1	*BTG2, CEBPB, DOK1, DSIPI, DUSP10, E2F2, ELF4, EVI1, FOS, FOSL1, HCK, HMGA1, HRAS, IL2RA, IL2RB, IL2RG, IL4R, IRS2, JUNB, LCK, LEF1, LTB, MADH3, MPL, NFKB2, NFKBIA, PLAU, RUNX1, SOX4, STAT5A, STAT5B, TP53, TRA1, ZFHX1B, ZNF145*	Tissue Morphology (n=25) Cellular Growth and Proliferation (n=28) Cellular Development(n=27)
Network 2	*CALD1, CCND2, CCND3, CTNNA1, ETS1, FLI1, HES1, MYB, MYC, MYCN, NFATC1, NOTCH1, NOTCH2, PAX5, PIM1, PRDM1, PRDX2*	Cellular Development (n=13) Hematological System Development and Function (n=10) Cancer (n=12)
Network 3	*BCL11A, CAPG,CCL4, CCL5, IFNGR2, IL6ST, INPP5A, KIT, MAP4K2, PTP4A3, PTPRE, PXN, SOCS2, SWAP70*	Hematological System Development and Function (n=22) Cell Death (n=23) Immune Response (n=21)
Network 4	*C3AR1,EPS15, HHEX, IFI30, MEF2C, MEF2D, NCOR1, NP, RXRA, ST13, TIE, ZFP36*	Gene Expression (n=17) Cellular Development (n=12) Cancer (n=13)
Network 5	*BCOR, CCND3, E2F2, GFI1, HOXA9, LMO2, MEIS1, TWIST1*	Cancer (n=21) Gene Expression (n=22) Cellular Growth and Proliferation (n=22)

**Complete networks are shown in supplementary Figures 1 – 5.*

Table 2. VIS/SAM genes in regulatory networks.

5.4.5 Networks based on VIS genes

We imported all VIS/CIS genes from Gr-1.4, BXH2 and AKxD MuLV models that were differentially expressed in the adult AML panel into the Ingenuity application to place them in regulatory networks. From this list (n=125), 110 genes present in the IPKB (focus genes) were used for the generation of networks. Five highly significant networks, associated with cell growth and proliferation, hematopoietic cell development, cell cycle, and gene expression were identified (Table 2, Supplementary Figures 1-5). Network 1 existed exclusively of focus genes (n=35), suggesting that genes within this network are commonly deregulated in AML. Multiple genes in this network, i.e. *IL2RG*, *STAT5A*, *STAT5B*, *IL4R*, *HCK* and *IRS2* are involved in cytokine signaling. The *SOX4* gene encodes a transcriptional

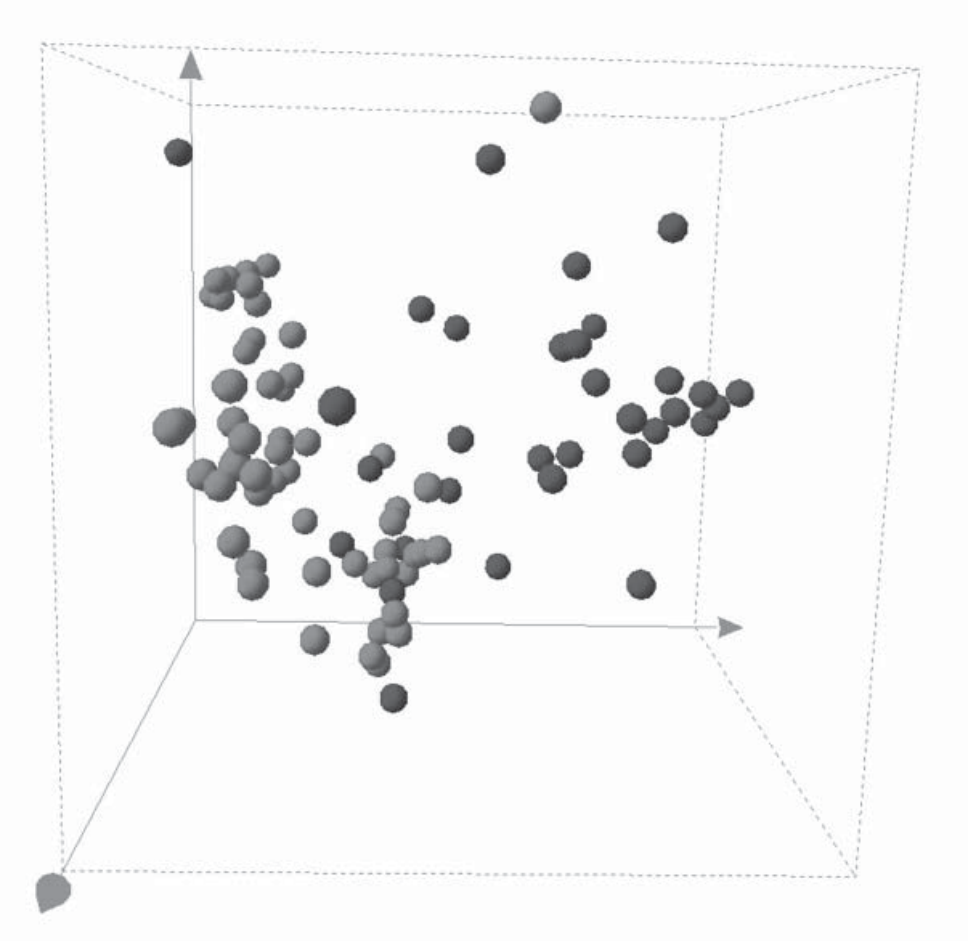


Figure 2. Principal component analysis showing clustering of AML patients, based on their expression signature of genes in network 5. *A comparison is shown of cases from good cytogenetic risk categories (red symbols) versus cases from poor cytogenetic risk categories (blue symbols).*
A full-color version of this figure is provided on the CD.

regulator implicated in the pathogenesis of neuronal tumors and lymphoma (13, 14) *ZNF145*, which is involved in t(11;17) in acute promyelocytic leukemia (APL), encodes a transcriptional repressor also known as promyelocytic leukemia zinc finger (PLZF) that has recently been implicated as a regulator of stem cell renewal (15, 16). We also asked whether networks might be differentially affected in prognostic subgroups of AML. To this end, we applied principal component analysis (PCA), by which AML samples are clustered in a three-dimensional space based on expression correlations of genes of each of the separate networks. Thus far, only network 5 clearly discriminated between AML patients with favorable and unfavorable cytogenetic risk indication (Fig. 2). SAM analysis indicated that this distinction is predominantly based on differential expression of *HOXA9*, *MEIS1* and *CCND3*, which are upregulated in the unfavorable group and *BCOR* and *GFI1*, which are downregulated in the unfavorable group (Supplementary Tables 6a and 6b).

5.5 Discussion

Genes commonly flanking MuLV provirus integration sites in murine leukemia and lymphoma are generally considered disease genes (12), although this idea has recently been challenged (7). Moreover, retroviruses may affect gene expression over several hundreds of Kb, which makes assignment of the relevant target gene ambiguous (7). We have systematically compared different groups of potential target genes, located within, near or more distantly from the insertion site with differentially expressed genes in subtypes of human AML, classified based on gene expression profiles. Our key finding is that genes located in direct proximity of the virus integration have a significantly higher probability to contribute to the gene expression-based clustering of both pediatric and adult AML than random genes, or than genes located more distantly from the site of integration. The data thus suggest that genes directly flanking MuLV integrations are most suspicious for their involvement in disease, although they do not preclude that in some instances deregulation of more distant genes may contribute to leukemic cell growth. Conceivably, in extended screenings, a significant proportion of such genes would also be found as VIS or CIS genes.

Thus far, only about 50% of VIS genes were differentially expressed in subsets of human adult AML classified by GEP (10). This may have multiple, not mutually exclusive, reasons. First, because the subsets of AML were identified by unsupervised clustering analysis based on gene expression relative to the mean of all samples (10), some disease genes may not be recognized because they are deregulated in samples that are not clustered with this approach. This may be addressed by extending GEP on more patients, which may allow definition of additional patient clusters. Secondly, a virus-flanking gene may be involved in murine, but not human AML. This may apply to genes encoding transcription factors that activate promoter and enhancer elements in the virus LTR (17, 18). Finally, some genes identified in mice may not be deregulated in human AML at the transcriptional but at the translational/posttranslational level, or may be functionally altered due to mutations.

Consistent with previous molecular and cytogenetic studies, the networks affected in AML mainly comprise signaling molecules and transcription regulators involved in growth factor-controlled cell proliferation and survival and the transcriptional

control of myeloid differentiation (19). However, Gr1.4 VIS genes deregulated in AML also include genes involved in other mechanisms (Table 2 and supplementary Tables 3a and 3b). For instance, *TXNIP* and *PRDX2* act in cellular responses to oxidative stress, whereas *CTNNA1* has been implicated in cell differentiation. *CTNNA1* is a candidate tumor suppressor gene located at chromosome 5q3.1 in a region that is frequently deleted in myelodysplasia and AML (20).
An important implication of this work is that disease genes and non-pathogenic genes, e.g., related to differentiation status of the cells, may be distinguished in clinical AML data sets. With the VIS gene lists in the various mouse leukemia models not yet saturated and the possibilities of GEP of AML still growing, the power of this strategy may increase. This may allow further refinement of currently identified and presumably disclose additional pathogenetic networks underlying AML. Such information would be useful for further refinement of diagnosis and for identification of key targets for therapeutic intervention.

References

1. Jonkers J, Berns A. Retroviral insertional mutagenesis as a strategy to identify cancer genes. Biochim Biophys Acta 1996, 1287(1):29-57.
2. Suzuki T, Shen H, Akagi K, et al. New genes involved in cancer identified by retroviral tagging. Nat Genet 2002, 32(1).166-174.
3. Erkeland SJ, 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.
4. Joosten M, Vankan-Berkhoudt Y, Tas M, et al. Large-scale identification of novel potential disease loci in mouse leukemia applying an improved strategy for cloning common virus integration sites. Oncogene 2002, 21(47):7247-7255.
5. Mikkers H, Allen J, Knipscheer P, Romeijn L, Hart A, Vink E, Berns A. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nat Genet 2002, 32(1):153-159.
6. Li J, Shen H, Himmel KL, Dupuy AJ, et al. Leukaemia disease genes. large-scale cloning and pathway predictions. Nat Genet 1999, 23(3):348-353.
7. Neil JC, Cameron ER. Retroviral insertion sites and cancer: fountain of all knowledge? Cancer Cell 2002, 2(4):253-255.
8. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 2004, 350(16):1605-1616.
9. Ross ME, Mahfouz R, Onciu M, et al. Gene expression profiling of pediatric acute myelogenous leukemia. Blood 2004, 104(12):3679-3687.
10. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 2004, 350(16):1617-1628.
11. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 2001, 98(9):5116-5121.
12. Akagi K, Suzuki T, Stephens RM, Jenkins NA, Copeland NG. RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res 2004, 32(Database issue):D523-527.
13. Lee CJ, Appleby VJ, Orme AT, Chan WI, Scotting PJ. Differential expression of SOX4 and SOX11 in medulloblastoma. J Neurooncol 2002, 57(3):201-214.
14. Shin MS, Fredrickson TN, Hartley JW, Suzuki T, Agaki K, Morse HC, 3rd. High-throughput retroviral tagging for identification of genes involved in initiation and progression of mouse splenic marginal zone lymphomas. Cancer Res 2004, 64(13):4419-4427.
15. Buaas FW, Kirsh AL, Sharma M, et al. Plzf is required in adult male germ cells for stem cell self-renewal. Nat Genet 2004, 36(6):647-652.
16. Costoya JA, Hobbs RM, Barna M, et al. Essential role of Plzf in maintenance of spermatogonial stem cells. Nat Genet 2004, 36(6):653-659.
17. Barat C, Rassart E. Members of the GATA family of transcription factors bind to the U3 region of Cas-Br-E and graffi retroviruses and transactivate their expression. J Virol 1998, 72(7):5579-

5588.
18. Barat C, Rassart E. Nuclear factors that bind to the U3 region of two murine myeloid leukemia-inducing retroviruses, Cas-Br-E and Graffi. Virology 1998, 252(1):82-95.
19. Lowenberg B. Prognostic factors in acute myeloid leukaemia. Best Pract Res Clin Haematol 2001, 14(1):65-75.
20. Horrigan SK, Arbieva ZH, Xie HY, et al. Delineation of a minimal interval and identification of 9 candidates for a tumor suppressor gene in malignant myeloid disorders on 5q31. Blood 2000, 95(7):2372-2377.

Chapter 6

HeatMapper: Powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics

Roel G.W. Verhaak[1], Mathijs A. Sanders[1], Maarten A. Bijl[1], Ruud Delwel[1], Sebastiaan Horsman[2], Michael J. Moorhouse[2], Peter J. van der Spek[2], Bob Löwenberg[1], Peter J.M. Valk[1]

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

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

BMC Bioinformatics. 2006 July 12;7(1):337.

6.1 Abstract

6.1.1 Background

Accurate interpretation of data obtained by unsupervised analysis of large scale expression profiling studies is currently frequently performed by visually combining sample-gene heatmaps and sample characteristics. This method is not optimal for comparing individual samples or groups of samples. Here, we describe an approach to visually integrate the results of unsupervised and supervised cluster analysis using a correlation plot and additional sample metadata.

6.1.2 Results

We have developed a tool called the HeatMapper that provides such visualizations in a dynamic and flexible manner and is available from http://www.erasmusmc.nl/hematologie/heatmapper/.

6.1.3 Conclusions

The HeatMapper allows an accessible and comprehensive visualization of the results of gene expression profiling and cluster analysis.

6.2 Background

Gene expression profiling by applying microarrays followed by cluster analyses is a powerful way to define pathobiologically relevant relations between the expression of sets of genes and disease classes. Unsupervised methods such as cluster analysis (1) and principal component analysis (2) are often applied to calculate and visualize these relations. Interpretation of results obtained by cluster analysis is frequently performed by visual inspection of a so-called heatmap; a matrix of genes versus samples in which gene expression levels or ratios are indicated using colors. Green often indicates low expression or down-regulation while red is frequently used to indicate high expression or up-regulation of genes (1, 3). A dendrogram, which is typically produced by unsupervised cluster analysis, provides further insights into sample-to-sample or gene-to-gene relations (1). These visualizations are useful when small numbers of samples and genes are analyzed, but are insufficient when studying larger datasets. Similarities and differences between samples or genes are easily lost due to the large size of these visualizations. This shortcoming particularly affects patient-cohort studies, since these analyses include increasing numbers of samples to allow comprehensive analyses.

A second type of heatmap that is frequently used is a matrix of pair-wise sample correlations in which anti-correlation or correlation is indicated by a color-scale, e.g. blue to red (4-6). Although details on individual gene expression measurements are lost, similarity between any pair of samples can easily be inspected.

To be able to correctly interpret both the sample versus gene expression heatmap and the sample versus sample correlation plot, data of the type of samples profiled, e.g. clinical parameters, karyotypes, mutations in particular genes, or gene expression data should be available. This information might then be included in a visual overview, as is frequently seen with sample versus gene heatmaps (7, 8). Such presentation would be a useful addition to the sample-sample heatmaps, which are frequently shown without metadata. Here we developed a tool, called the

HeatMapper, which can generate such combined visualizations. The tool is simple in use and allows dynamic and flexible display of a correlation plot in combination with sample characteristics.

6.3 Implementation

The HeatMapper, written in JAVA (version 1.4.2), uses comma-separated or tab-delimited text-files as input. It requires two files: one file containing a matrix of sample-sample similarity, i.e. Pearson correlation, Spearman correlation or Euclidean distance, and one file with sample related data. In both files, similar sample ID's are used. Correlation files can be generated using tools such as Omniviz, GeneMaths and R/BioConductor, while sample data files can for instance be created in Microsoft Excel. Example files are available from the website. Alternatively, the tool can be adapted to communicate with a database. In our laboratory, the HeatMapper is connected to a MySQL database which further optimizes the workflow. This version is available on request.

6.4 Results & Discussion

As the upper right part of a traditional sample versus sample heatmap is in fact a mirror image of the lower left part, it is redundant. Therefore, when data are loaded, the HeatMapper only displays a triangular heatmap (Figure 1). Sample-sample (dis-) similarity, i.e. Pearson correlation, Spearman correlation or Euclidean distance, is mapped to a color scale ranging from blue to red. Dark blue relates to the negative extreme value of the metric, i.e. -1 for Pearon correlation, where dark red refers to the positive extreme value, i.e. 1 for Pearson correlation. Sample related data, can be simply added via the menu and is subsequently plotted alongside the heatmap diagonal. Different entries in one sample characteristic are mapped to different colors, or, in the case of numeric data, shown as bars of which the size is proportional to the value. Several options are available to customize the resulting visualization, such as zoom functionality and options to change the colors used in histograms or bars to indicate phenotypic or genotypic differences. Further customization options include the possibility to change the sample order, allowing a user for instance to visualize the results of a different clustering algorithm, or to sort the data according to any user-defined order. This can be accomplished via selecting the 'Change sample order' menu-option, after which the order of the sample ids can be inserted by typing them or using copy-paste. Subsets of the original data can be created and viewed in any sequence. Importantly, high-resolution images of the produced figures can be exported using the Portable Network Graphics (PNG) format.

Our tool provides several advantages over more traditional means of presenting results obtained gene expression profiling and clustering analysis (7, 8). The pair-wise display of samples clearly indicates similarity in expression profiles. By combined visualization of sample versus sample similarities and sample characteristics, subclasses of samples sharing a commonality, such as a mutation in a particular gene, and a high similarity in expression profile can be readily identified. Cluster assignments, made manually by the user, can then be added via the 'Add special values' menu option and displayed as sample characteristic.

As an example, Figure 1 shows the results of a cluster analysis of 285 acute myeloid

Figure 1 (facing page). HeatMapper screenshot. *The figure shows pairwise correlations between 285 samples of patients with Acute Myeloid Leukemia, as described previously (6). The cells in the visualization are colored by Pearson's correlation coefficient values with deeper colors indicating higher positive (red) or negative (blue) correlations. Clinical and molecular data are depicted in the columns along the original diagonal of the heatmap. Karyotype and FAB classification based on cytogenetics are depicted in the first two columns (karyotype: normal-green, inv(16)-yellow, t(8;21)-purple, t(15;17)-orange, 11q23 abnormalities-blue, 7(q) abnormalities-red, +8-pink, complex-black, other-gray; FAB M0-red, M1-green, M2-purple, M3-orange, M4-yellow, M5-blue, M6-grey). FLT3 ITD, CEBPa and NPM1 mutations are depicted in the same set of columns (red bar: positive and green bar: negative). The expression levels of CD34 (probe set: 209543_s_at) in the 285 AML patients are plotted in the last column (bars are proportional to the level of expression).*
A full-color version of this figure is provided on the CD.

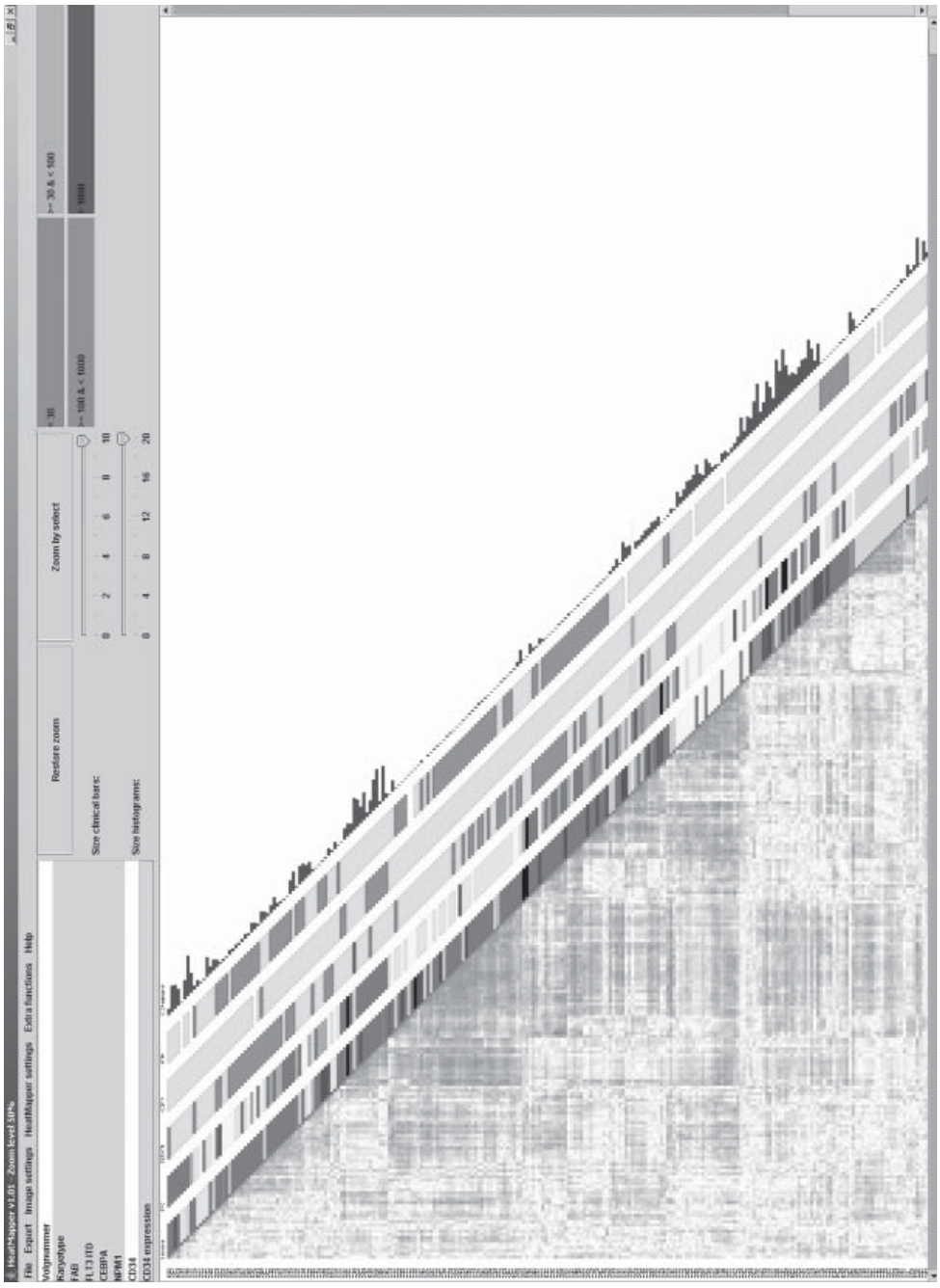
leukemia (AML) samples. Clusters are recognized as red triangles near the plot diagonal. Sample related data are presented in the adjacent bars, where the same color indicates the same characteristic. The last bar indicates the expression levels of CD34, in which the level of expression is proportional to the length of the bar. By visual inspection of this plot, one can immediately conclude that (1)AML samples can be separated into several subtypes, such as cases with a t(8;21), based on expression profiling (9), (2) several clusters are related to a single distinguished abnormality (for instance nucleophosmin (*NPM1*) mutations), indicated in red in the fifth column and (3) mRNA levels of CD34 are low in samples with *NPM1* mutations.
In our laboratory the HeatMapper code has been coupled to a database containing gene expression profiling results, from which gene expression levels can dynamically be obtained. This allows the quick and accurate visual inspection of the distribution of expression levels in different clusters, and making the tool even more powerful. The database implementation, is available on request.
Our visualization method has been successfully applied in several studies (6, 9-12).

6.5 Conclusion

With the increase of the number of samples profiled, particularly in patient-cohort studies, specialized visualization methods for microarray studies are indispensable. Our tool allows the accurate inspection of combinations of dataset characteristics, i.e. correlations and clustering results and sample related characteristics, i.e. survival time and gene expression levels. Summarizing, the HeatMapper tool results in powerful visualization tool that allows the accurate and rapid interpretation of the data obtained by large scale gene expression profiling. The HeatMapper tool has already proven to be very useful in several studies (6, 9-12).

Availability & Requirements

Project name: HeatMapper
Project homepage: http://www.erasmusmc.nl/hematologie/heatmapper/
Operating system: Platform independent
Programming language: JAVA


HeatMapper v3.01 - Zoom level 50%
File Export Image settings HeatMapper settings Extra functions Help
Volgnummer
Karyotype
FAB
FLT3ITD
CEBPA
NPM1
CD34
CD34 expression
Restore zoom
Zoom by select
Size clinical bars:
Size histograms:
< 30
>= 100 & < 1000
>= 30 & < 100

Other requirements: JAVA 1.4.2 or higher.
License: The tool is available free of charge. Source code is available upon request.
Any restrictions to use by non-academics: None

List of Abbreviations

AML	Acute Myeloid Leukemia
PNG	Portable Network Graphics
NPM1	Nucleophosmin

Author's contributions

RGWV designed the software, participated in all phases of research and wrote the manuscript; MAS wrote the majority of the JAVA code; MAB contributed to software design and earlier code; RD gave intellectual contributions and revised the manuscript; SB contributed to the software code; MJM and PJS were involved in an earlier implementation of the software; BL gave intellectual contributions; PJV initiated the idea, gave intellectual contributions and revised the manuscript.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998, 95(25):14863-14868.
2. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000:455-466.
3. Spellman PT, Sherlock G, Zhang MQ,et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998, 9(12):3273-3297.
4. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 2002, 97(457):77-87.
5. Ross ME, Mahfouz R, Onciu M, et al. Gene expression profiling of pediatric acute myelogenous leukemia. Blood 2004, 104(12):3679-3687.
6. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 2004, 350(16):1617-1628.
7. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002, 415(6871):530-536.
8. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005, 365(9460):671-679.
9. Bullinger L, Valk PJ. Gene expression profiling in acute myeloid leukemia. J Clin Oncol 2005, 23(26):6296-6305.
10. Valk PJ, Delwel R, Lowenberg B. Gene expression profiling in acute myeloid leukemia. Curr Opin Hematol 2005, 12(1):76-81.
11. van den Akker E, Vankan-Berkhoudt Y, Valk PJ, Lowenberg B, Delwel R. The common viral insertion site Evi12 is located in the 5'-noncoding region of Gnn, a novel gene with enhanced expression in two subclasses of human acute myeloid leukemia. J Virol 2005, 79(9):5249-5258.
12. Verhaak RG, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. Blood 2005, 106(12):3747-3754.

Chapter 7

Discussion

7.1 *Diagnostics of acute myeloid leukemia through gene expression profiling*

Acute myeloid leukemia (AML) is a heterogeneous disease from various points of view. Leukemia develops through a series of accumulating genetic changes. As a result, the clinical phenotype of leukemia is diverse. This has an impact on the development, the choice and management of therapy of patients with AML. This thesis dealt with various aspects of genome-wide expression profiling, an approach that holds promise for developmental therapeutics in cancer in general and leukemia in particular.

7.1.1 Recognition and prediction of known AML subtypes using gene expression profiling

Recently, gene expression profiling using microarrays has become a popular method to identify and predict known prognostically and clinically relevant subclasses in cancer (1-3). A well-known example is the study of Golub et al, in which AML and ALL samples could be distinguished based on their expression profiles (1). In another more recent study, a large number of 937 samples, including chronic and acute lymphoid and myeloid leukemia, and normal controls, were analyzed using gene expression profiling (4). This study dealt with establishing the diagnosis of global categories of acute and chronic leukemia using expression profiling and confirmed that apart from other hematologic malignancies, acute and chronic myeloid and lymphoid leukemia can be distinguished.

Another challenge is to verify whether expression profiling can also be applied to recognize distinct subtypes within a particular hematologic malignancy, like AML. In recent years, it has become evident that microarrays are suitable for distinguishing certain subclasses of AML from others. For instance, AML with a trisomy of chromosome 8 could be distinguished from samples with normal cytogenetics (n=27) (5); cases of AML with mixed leukemia lineage gene (*MLL*) rearrangements could be discerned from both other ALL and AML cases (n=65) (6). Samples with reciprocal translocations (inv(16), t(15;17) and t(8;21)) could be distinguished from other AML samples (n=37) (7), and AML with *MLL*- and reciprocal rearrangements could be separated from AML carrying normal karyotypes (n=28) (8).

These conclusions were derived in studies including cases with a limited spectrum of subtypes of AML. In Chapter 2, we used a broad, more representative cross-section of AML and analyzed cellular gene expression profiles of a patient-cohort of 285 patients, i.e., without excluding any particular subtype of AML, and investigated whether known subtypes could be identified (Chapter 2). We found that recurrent reciprocal translocations such as inv(16), t(15;17) and t(8;21) each have highly distinct expression profiles, while *MLL* translocations aggregated in two separate clusters. Thus, it appears that certain common karyotypic aberrations can indeed be recognized using gene expression profiling.

During the last decade many novel small molecular aberrations have been identified (Table 1B, Chapter 1). Several efforts have been made to derive their characteristic gene expression signatures. It was found that the expression signature of AML with *FLT3* internal tandem duplication and *FLT3* tyrosine kinase point mutations can be predicted with high accuracy (9, 10). On the other hand, AML with N-*RAS* mutations could not be predicted according to a distinctive expression signature

in studies by us and others (10). In our study, *FLT3* internal tandem duplication mutations were correctly predicted in 79% of samples in a validation dataset. In contrast, we were unable to predict FLT3 tyrosine kinase mutations through a specific expression pattern in our study. The discrepancy between the latter results with those of Neben et al (10) could be due to the greater numbers of samples with tyrosine kinase mutations in the Neben et al (10) study, allowing the detection of a more subtle signature. Other molecular abnormalities were found to be associated with specific gene expression signatures in our study, such as *CEBPa* mutations, which segregated into two separate gene expression groups with different cytological FAB-types. Likewise, AML cases with *NPM1* mutations (Chapter 4) clustered in various subgroups, some of which co-associated with the presence of *FLT3* internal tandem duplication mutations. AML cases with high *EVI1* expression were also found to be associated with particular gene expression patterns. Thus, molecular abnormalities have pronounced effects on genome-wide mRNA expression, although not always as prominently as balanced cytogenetic translocations such as inv(16), t(15;17) and t(8;21).

Comparing several malignancies including AML, Haferlach et al (4) recently showed that six AML subgroups, respectively with inv(16), t(15;17), t(8;21), complex karyotypes, 11q23 and normal karyotypes/other, were predictable through gene expression profiling with sensitivities and specificities comparable to our results (***Chapter* 2)**. The results of these representative studies show that gene expression profiling can be used to identify distinct molecular subtypes of AML, such as AML with inv(16), t(15;17), t(8;21), mutant *CEBPa* or high *EVI1* expression. Several other AML subclasses, i.e. those with N-*RAS,* K-*RAS,* trisomy(8), *FLT3* tyrosine kinase point mutations, do not relate with a distinct signature. We assume that the latter AML types are too heterogeneous to be predicted by means of gene expression profiling, e.g. due to effects of additional cooperative mutations with stronger signatures.

7.1.2 Identification of unknown subtypes of AML using gene expression profiling

Gene expression profiling in AML has further highlighted the extraordinarily heterogeneous nature of the disease (4-7, 9-11). Although the majority of leukemia cases are associated with one or more genetic abnormalities, not all of these abnormalities are prognostically useful (Table 2, Chapter 1). In fact, 15-20% of patients lack any prognostically relevant molecular abnormality. Gene expression profiling might be useful for the identification of novel subtypes of AML with prognostic variations. Currently, the number of studies dealing with this issue is still limited. Using semi-supervised clustering of 116 samples, profiled on Stanford dual-channel microarrays, Bullinger et al. were able to distinguish patients with a normal karyotype with a good from those with an adverse prognosis, based on a signature of 133 genes (11). Their signature was recently validated in an independent dataset (n=64) (12). Several factors, such as mutations in *FLT3* and particular FAB subtypes, were unevenly distributed between the two groups. Multivariate analysis, for which a larger number of samples is required, will have to show whether the signature also harbors independent prognostic value. In our study, we identified a subset of AML cases (cluster #10) with a common gene

expression signature that was associated with a variety of known poor prognostic markers, such as high *EVI1* expression, t(9;22) and monosomies 5 and 7 (***Chapter 2***). Using a training- and a validation-set, we found that this cluster #10 subtype of AML could be predicted with high accuracy. Interestingly, the gene expression profile of "cluster #10 AML" resembles the gene expression signature of normal CD34-positive hematopoietic precursor cells. Recently, Heuser et al showed that poor responders to induction chemotherapy also cluster together with normal CD34-positive cells, indicating a possible overlap between these two groups (13). The co-clustering of "cluster #10 AML" and CD34-positive cells might indicate that the phenotypes of these types of leukemia resemble early hematopoietic progenitors and are less responsive to therapy.
The "cluster #10 AML" represents a new subgroup of AML with a poor prognosis, that is of similar frequency as, for instance, the t(8;21) or inv(16) subgroups. Identification of new AML subtypes using microarrays deserves further investigations, e.g. when extended series of cases are examined.

7.1.3 Prediction of prognosis

An important goal of gene expression profiling research in AML is to establish a individualized classification into good and poor prognosis. This is as yet not possible, due to insufficient resolution of the heterogeneity of AML. In an effort to derive prognostic information from gene expression patterns, we have also tried to classify our cohort into AML with poor or good prognosis, by comparing the profiles of patients in continuous complete remission with those of patients with a relapse of leukemia. The result of this analysis revealed an unacceptable misclassification error of 40%, using a training- and a validation dataset. A similarly negative result was obtained when the same experiment was restricted to cases of AML with a normal karyotype only. It appears that AML is too heterogeneous to accurately predict prognosis with only one classifier. Several clusters show up with a clear gene expression profile but do not correlate with distinctive survival characteristics. The number of cases of some of these clusters was too small for reliable survival estimates.
We foresee that in the future, useful prognostic classifiers will be defined within particular genotypic subtypes of AML. Indeed, within AML with t(15;17), a subset with a *FLT3* ITD and concordantly a high white blood cell count, appear to cluster separately (Figure 1, Chapter 2). *FLT3* ITD and high white blood cell count in APL with t(15;17) tend to have a comparatively worse prognosis (14).

7.1.4 Aggregation of microarray data

To increase the number of samples, efforts have been made to aggregate data from various expression profiling studies to increase the number of samples. We used Affymetrix HG-U133A GeneChips whereas Bullinger et al (11), used a different biochip platform, i.e. Stanford dual channel arrays. The limited overlap of genes of these two types of arrays prohibits a thorough concurrent analysis. In another profiling study, the investigators used the same Affymetrix HG-U133A GeneChip that we have employed (4). Since the data of the latter study have not been released in the public domain, a comparative analysis or meta-analysis has not been directly possible. In a third study, pediatric AML was profiled on the same Affymetrix

HG-U133A microarray platform (15). In this case, we have taken advantage of the opportunity and analyzed their and our 415 adult and pediatric samples using the same bioinformatics approach in direct comparison (Figure 1). Subtypes with specific chromosomal abnormalities such as inv(16), t(8;21) or t(15;17) in both the pediatric and adult series revealed distinct and similar gene expression profiles. It is of note that while pediatric and adult samples aggregated together, they formed adjacent subclusters within each cluster. It is unclear whether this small difference between the pediatric and adult AML genotypes is due to the age differences of both cohorts, or differences in processing of samples in the two laboratories.

7.1.5 Perspectives of expression profiling in AML

Our results demonstrate that the clustering of the AML cases based on gene expression profiling is driven by two phenomena: (I) the common karyotypic or molecular abnormalities, such as balanced translocations, and (II) differences in cytological maturation status of the primary AML. At this point, particular genetic subgroups, e.g. AML with internal tandem duplications in *FLT3*, cannot yet be identified at the accuracy level required for clinical applications. By including genes currently not present on the arrays, or prediction in larger subsets of AML patients this problem may potentially be overcome. Analyzing such datasets will reveal not only additional, but also more specific genes that relate to particular AML subtypes.

By selecting the distinctive and informative probe sets, a specialized microarray for AML diagnostics can be developed. Subsequently specialized GeneChips can be used for proper diagnosis of AML using one assay. Interestingly, such microarrays have been developed for lymphoid leukemia and breast cancer (16, 17).

7.2 Technical improvements of oligonucleotide array research

Until recently, microarrays did not represent the whole genome but a fraction of 40-45% of the genes. It is likely that genes critical in the leukemogenic process are missed. With further technological advancements, for instance due to further shrinking of feature size, a whole genome microarray is foreseen to become available in the future. Currently, commercial platforms such as Affymetrix represent 60-65% of the genome. In addition, the sensitivity of microarrays, which are currently less efficient in measuring low expression values (18), is also likely to improve. The quality of probe set design can probably be increased through the use of more accurate information in sequence databases. Better understanding of alternative splicing of mRNA transcripts and higher coverage of alternative splice variants on microarrays will further contribute to an increased insight into cellular processes. Affymetrix exon arrays, containing probe sets for exons rather than genes, are currently being tested (54).

From a statistical viewpoint, other methods to estimate noise and background- and cross hybridization, such as quantile normalization and variation stabilizing normalization (19, 20), have been developed. Similarly, alternative procedures for estimating expression levels, such as dChip and robust multi-array average (RMA) were demonstrated (21, 22). Affymetrix has recently released a new algorithm designated PLIER, which applies a similar approach as RMA to measure expression (23). However, although the new algorithms have advantages, no single algorithm

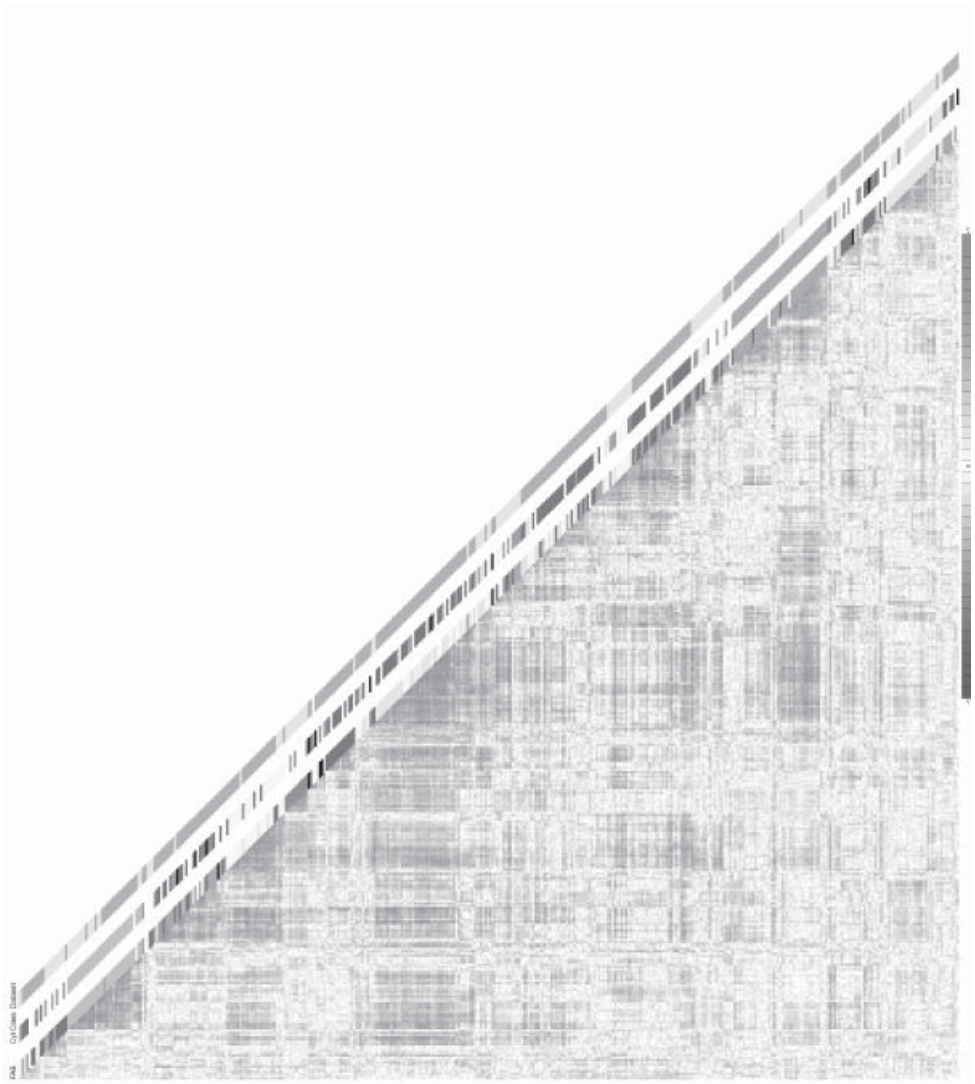


Figure 1. Unsupervised cluster analyses of 130 pediatric (15) and 285 adult (Chapter 2) acute myeloid leukemia (AML) patients . *Patients were clustered using 2,175 Affymetrix probe sets which are representative for the 2,856 probe sets used for the most optimal clustering of the adult AML with regard to known molecular abnormalities. French-American-British (FAB) classification and karyotype based on cytogenetics are depicted in the first two columns along the original diagonal of the Correlation View (FAB M0, red; M1, green; M2, purple; M3,orange; M4, yellow; M5, blue; M6, gray; M7, red. Karyotype: normal, green; inv(16), yellow; t(8;21), purple; t(15;17), orange; t(11q23)/MLL abnormalities. Blue; 7(q) abnormalities, red; trisomy8, pink; complex, black; other, gray). In the third column, pediatric AML cases are indicated in green and adult AML cases in red.* ***A full-color version of this figure is provided on the CD.***

has yet shown to be fully able to distinguish between noise and signal under all circumstances (24). We showed that the outcome of analysis of microarray data is influenced by the choice of the pre-processing method. This effect is most obvious as demonstrated by the analysis at the probe set level, i.e. individual expression levels and identification of differentially expressed genes (Chapter 3). As the data compared in our study could not be compared with data representing the ground truth, the "golden standard" pre-processing method could not be identified.
With further improvements as regards for instance technical design of oligonucleotide arrays, sophisticated algorithms for normalization and expression measurement adapted to size and type of study, closer estimates of true gene expression will be obtained with oligonucleotide arrays.
In another approach for improving analysis outcomes in large datasets, the variability between probes within a probe set is tested and probe sets are disqualified if variation exceeds a certain threshold. The quality of the dataset on which all further analyses are performed can thus be improved, as less effective probe sets are omitted. This approach, which is called factor analysis, is still largely experimental (25).

7.3 AML pathogenesis

The large quantities of data produced in microarray experiments may also be useful for elucidating the molecular pathogenesis of leukemia. We have proposed a novel approach towards classifying samples based on pathogenesis, by combining genes present in the signature of the 16 clusters of clinical leukemia with sets of leukemia genes identified in retroviral insertional mutagenesis experiments in experimental animals (Chapter 5). We found that genes found close to virus integrations sites are more frequently differentially expressed in human AML, supporting their involvement in leukemogenesis.
Using retroviral insertional mutagenesis, predominantly activated genes, i.e. proto-oncogenes, have been detected. Retroviral insertional mutagenesis may besides activating genes also repress genes that play a role in leukemogenesis. The fact that integration of virus may be followed by methylation of the integration site may provide a key to the identification of repressed genes (26). Experiments that take advantage of the methylation of integration sites result in a higher detection rate of genes involved in murine leukemogenesis. Furthermore, more accurate gene lists differentially expressed in AML will be obtained due to advances in microarray research, e.g. larger datasets and microarrays containing more genes. Combined with the increasing possibilities of pathway analysis software, increasingly accurate and complete biochemical interaction networks will be constructed. As these networks represent the candidate pathogenetic networks involved in AML, these advances potentially allow a classification of AML based on pathogenesis.
Genes that are differentially expressed in a particular cluster or class of AML can either be disease genes, i.e. involved in the development of the disease, or just downstream marker genes. Bioinformatics tools, such as Toucan (27, 28), have been developed to detect statistical overrepresentation of transcription factor binding sites in the promoter regions of a list of genes. By applying the Toucan software, transcription factor target genes were identified (29). To identify direct target genes of transcription factors involved in AML, we have applied this methodology on lists

of genes differentially regulated in AML. Toucan was employed on the expression signatures of clusters with core binding factor complex abnormalities, i.e., cluster 9 (inv(16)) and cluster 13 (t(8;21)). The core binding transcription factor complex is constituted of several proteins, one of them being AML1. AML1 can bind specific DNA sequences, also called AML1 binding sites. Genes deregulated in core binding factor leukemia cases are potential targets of the core binding factor complex. Promoter sites of these targets are likely to contain one or more AML1 binding sites. However, using Toucan, we did not find any significant overrepresentation of AML1 binding sites in promoters of core binding factor target genes. The lack of AML1 binding sites could be due to the small size of the AML1 binding site (6 bp) resulting in high background levels. The fact that active binding sites may be present at large distance of a gene might contribute to this problem. Furthermore, it is possible that the lists of genes used as input contain downstream targets, which lack AML1 binding sites. The failure to find evidence for binding sites could also be due to the lack of knowledge of binding site sequences, or the amount of noise present in both lists of differentially expressed genes (genes that are associated with a particular subgroup but are not a target of one particular transcription factor) and promoter sequences (short binding site sequences randomly occurring due to chance). However, a list consisting of known AML1 target genes also failed to reveal AML1 binding sites that were statistically overrepresented.

Research aimed at elucidating the pathogenesis of AML may take advantage of data generated by expression profiling of clinical material. For instance, our gene expression profiling study (Chapter 2) revealed two clusters (cluster #4 and cluster #15) dominantly containing cases of AML with mutations in *CEBPa*. In fact, cluster 15 consists exclusively of *CEBPa* mutated AML samples, while cluster 4 contains 9 samples with and 6 samples without *CEBPa* mutations. Interestingly, AML cells with *CEBPa* mutations display high expression levels of *CEBPa* mRNA, while cell samples from AML cases without mutation do not show any measurable mRNA expresssion. A possible explanation for the lack of *CEBPa* mRNA expression may relate to promoter methylation. By assessing methylation status of the *CEBPa* promoter in all 15 samples we could indeed show that hypermethylation of the *CEBPa* promoter was only apparent in those AML cases with no *CEBPa* mutations and no measurable expression of *CEBPa* transcripts (30). Whether methylation is the cause or the result of mutated *CEBPa* remains subject of further research. Yet, this example of two gene expression signatures in relation to *CEBPa* mutations illustrates that gene expression profiling may open ways towards deconstructing the molecular basis of leukemogenesis.

Sophisticated bioinformatics approaches have been developed to derive pathogenetic relevant relations from microarray data. Several groups have tried to identify sets of genes that correlate in aggregated microarray datasets. Correlations between expression levels potentially indicate functional interaction of gene products (31-33). By compiling a database of signatures of different tumor- and tissue types, machine-learning techniques can be applied to find commonalities amongst the expression profiles. Sets or modules of genes that show comparable trends in different datasets are suggested to be functionally connected. Enlarged datasets and data available from other sources will contribute to a more complete understanding of biomolecular interactions (34, 35).

Data mining of scientific literature provides an interesting approach for gaining information on intracellular reactions. Lists of genes, for instance generated through microarray research, are used as input to retrieve information on interactions between genes, and potential biochemical networks are generated. Several commercial and academic approaches have been presented, such as Ingenuity® (36), MetaCore™ (37) and ACS (38). A limitation of this approach is that at the current state of development, knowledge about interactions is only in part available in literature or moreover, not all published data may be accurate. It is therefore not possible to confidently construct complete networks of cellular processes using this approach. These applications will gain in value along with ongoing development and increased understanding.

7.4 High-throughput techniques in cancer research

Originally, oligonucleotide arrays have been developed for high-throughput DNA and even peptide research (39-41). Although expression profiling is widely applied, DNA profiling is still an emerging field at an early stage of development. Further development of the early arrays for targeting DNA sequences have now resulted in the production of microarrays for probing single nucleotide polymorphisms (SNP), so-called SNP-arrays. Such arrays target 10.000 to 500.000 common SNP's, which are more or less equally divided over the genome. Not only genomic polymorphisms are established, but also DNA copy numbers. This technique can thus be used to get information on chromosomal gains or losses. The genotype calls provide data of loss of heterozygosity. An alternative approach to DNA profiling is array-based comparative genomic hybridization (array-CGH). As dual-channel microarrays, array CGH uses DNA from a test and a reference sample to assess copy numbers, but this technique cannot by applied to assess genotypes (42). Current resolutions are comparable to those of SNP-chips. Interestingly, SNP-array analyses of AML have revealed the presence of uniparental disomy, probably as a result of mitotic recombination, in about 20% of AML patients (43, 44).

Losses or gains of chromosomal materials detected with DNA-profiling may have effects on mRNA transcript levels and prognostically relevant chromosomal abnormalities could therefore also be identified through expression profiling. However, genome-wide SNP-array analyses in AML may facilitate the identification of common abnormalities underlying distinct expression clusters.

Several other high-throughput techniques are currently in development or have recently been applied. Chromatin immunoprecipitation (ChIP) is a technique to study protein-DNA interactions (45) to assess e.g. the binding of transcription factors to particular DNA regions or the epigenetic status of a gene and its regulatory regions. . To facilitate large-scale analyses, ChIP has been combined with microarray technology (46). This allows the identification of target sites for a certain protein on a genome-wide level. In AML, target genes of disrupted transcription factors such as the core binding factor complex (involved in inv(16) and t(8;21)) or the PML-RARα fusion protein (involved in t(15;17)) could as such be identified. ChIP-on-Chip profiling in combination with gene expression profiling could be useful fo elucidating regulatory networks involved in leukemogenesis (47).

Methylation of CpG-islands in the genome is frequently seen and is known to play a role in regulating gene expression (48). Methylation profiling of cancer

may reveal specific patterns of CpG island methylation resulting from clonal selection of cells with growth advantages, e.g. due to silencing of associated tumor suppressor genes (49). For instance, methylation of genes associated with *CEBPa* mutations, possibly through the upregulation of methyl transferase enzymes such as *DNMT3B*, has been demonstrated (30). It is thought that this epigenetic event plays a role in leukemogenesis. Therefore, methylation profiling of AML might be of use in identifying critical regulatory genes that are silenced as a result of molecular abnormalities.

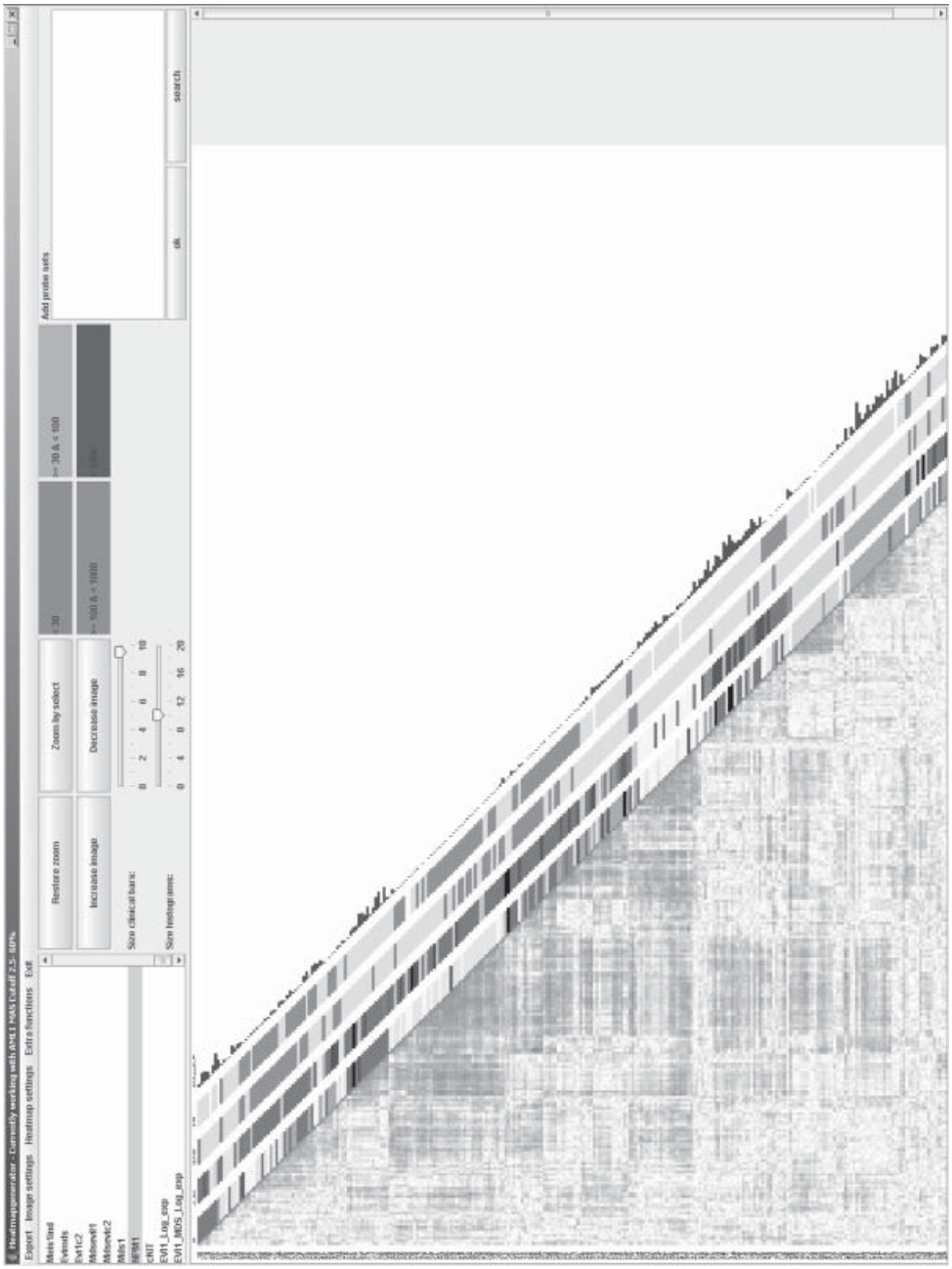
Alternative splicing of primary RNA transcripts results in different mRNA transcripts of the same gene and plays an important role in cell homeostasis (50-52). For instance, different splice forms of *EVI1* with different oncogenic characteristics have been shown to be present in AML (53). To identify the influence of alternative splicing, exon arrays have been developed targeting over 300.000 different Ensembl transcripts (54). Arrays targeting different splice variants can be used to identify particular active splice forms and specific combinations of co-occurring abnormalities that are involved in AML pathogenesis.

MicroRNA's are short length RNAs (22 bp) that represent a class of mRNA translation regulators (55). They have been postulated to have a role in the control of cell development, growth, maturation and other cellular processes. A single microRNA can have up to hundreds of target mRNAs, which are subject to mRNA degradation and translation repression. A variety of microRNAs have been identified in recent years. Profiling tumors for expression of a series of microRNAs has been shown to characterize and to distinguish different sorts of cancer (56). Profiling of microRNAs will in the future be applied to distinguish microRNAs that are associated with different subtypes of AML. Some of these microRNAs may be involved in pathogenesis of leukemia (57).

On the protein level, mass spectrometry is a promising technique for assessing protein content and protein levels in cells (58). However, this technique has to be further developed to reliably identify the proteome present in a biological sample. Currently, mass spectrometry reliably identifies only one to five percent of most abundant proteins present. Protein levels may correlate with mRNA levels but it is clear that there is no direct relationship between mRNA and protein levels, e.g. due to variations of cellular processing of mRNA (59). Mass spectrometry furnishes a more specific way of identifying the active components that play a role in the cellular processes in leukemia cells.

A key challenge in current high-throughput research is the identification of the most relevant targets for further investigation. Due to the large amounts of data

Figure 2 (facing page). Screenshot of the JAVA heatmap explorer, which is part of MADEx. *The right upper half of the (mirror image) heatmap is left out and several histograms, indicating different sample characteristics, have been added. In this example, the first bar indicates FAB status ; the second bar indicates karyotype. Different colors indicate different patient class. The third and fourth bar indicate presence of AML specific acquired mutations in the genes FLT3 and NPM1. Green indicates absence of a mutation while red indicates presence of a mutation. The fifth bar displays relative expression levels of the CD34 gene. All data is retrieved dynamically from the MADEx database.*
A full-color version of this figure is provided on the CD.


Export
Image settings
Heatmap settings
Extra functions
Exit
Restore zoom
Zoom by select
Increase image
Decrease image
Size clinical bars:
Size heatmaps:
Add probe sets
ok
search

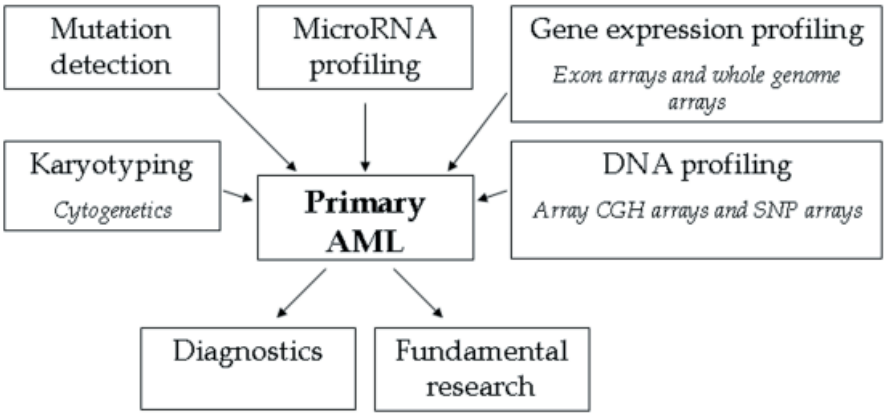


Figure 3. Applying multiple high-throughput techniques on a cohort of de novo AML samples will further improve diagnosis but will also be useful for fundamental research of AML.

involved, the number of false-positive and false-negative results will be significant even at low rates. It is therefore necessary that results are confirmed using other techniques. Combining data from different high-throughput sources, like SNP-arrays, methylation profiling and gene expression analyses, allows a technique-independent validation of results (Figure 3). As an example of the potential use of integrative genomic analyses, a prognostically relevant mutation in Microphthalmia-associated transcription factor was identified by comparing expression data with DNA copy number changes (60). From this example, we can learn that the complementary use of different techniques for complete characterization of patient samples can add to the understanding of tumorigenic and pathogenetic processes.

7.5 Managing large data volumes in cancer research

To be able to compare experiments and exchange data between applications, standardization of analysis and data is important. To this end, many journals require the publication of microarray data supporting research in a specific format called Microarray and Gene Expression Markup Language (MAGE-ML). It is an XML-format based on the Minimum Information About Microarray Experiments (MIAME) standard (61), developed by a consortium of academic and commercial research investigators called MGED (62). Several large public data-repositories such as ArrayExpress and the Gene Expression Omnibus accept MAGE-ML. The archiving of raw data has facilitated research of gene expression profiles on an aggregated level (31-33). Many academic and commercial efforts have been undertaken to develop database systems for storing and analyzing microarray experiments, such as Bioarray Software Environment (BASE), the Stanford Microarray Database (SMD) and the TIGR Microarray Suite 4 (TM4) (63-65). Also, applications have been developed that are available for local installation. These are used to analyze data, such as Cluster/Treeview and BRB ArrayTools (66, 67). An important contribution has been made through the development of the statistical

programming environment R (68), a open-source implementation of the S-plus statistical programming language (69). Together with Bioconductor, an extended library of predefined analysis methods, the R programming environment can be applied to analyze microarray data with any method required. A disadvantage of these applications is that they require a basic knowledge of data-analysis methods and that in general they are not easy in use. Furthermore, integrated visualization of large series of specimens in relation to various other sets of data, e.g. clinical parameters, other molecular data, is not available. In Chapter 6, we propose an approach to visualize sample-sample correlations in an integrated view with sample-specific characteristics, such as karyotype or survival. An implementation of this visualization has been made available in the HeatMapper, as described in Chapter 6.
The latter visualization has been implemented in the MicroArray Data Explorer (MADEx). In MADEx, we have combined the HeatMapper tool with a database application that stores expression data and analysis results from microarray experiments and allows access to these data via a user-friendly and flexible web-interface (Figure 2). By storing expression data together with results of different types of analysis, such as cluster analysis or tests for differential expression, MADEx functions as a central repository for microarray studies. The central storage of these types of data combined with several dynamic analysis- and visualization functionalities within MADEx, allows researchers to quickly access data of microarray experiments at different levels in a web-based manner. To date, MADEx has been successfully applied in several studies (70-73).

7.6 Conclusions

The investigations reported in this thesis furnish evidence as regards the potential utility of expression profiling using oligonucleotide arrays for the diagnosis of AML. Most likely, in the near future further improvements both in statistical methods and array technology will be introduced. Based on this, one may forecast that within several years expression profiling will replace, at least in part, certain current routine diagnostic approaches (immunophenotyping, cytogenetics, molecular diagnostics) and offer added informative value. The discovery of additional prognostically relevant mutations, such as mutations in *FLT3*, *CEBPa* and *NPM1*, will further contribute to the refinement of AML classification. Identification of such molecular changes will be greatly enhanced by applying high-throughput techniques that have recently become available, such as SNP profiling, microRNA profiling and methylation site profiling (Figure 3) .
Moreover, integration of data available from different data sources will allow for in depth analysis and contribute to the elucidation of the molecular pathogenesis of AML. As we showed, microarray data, in combination with data from murine virus integration models, can already be used to generate pathogenetic networks that define different classes of myeloid leukemia. Future research will result in extended list of murine AML genes, more complete sets of differentially expressed genes and more accurate definitions of protein-protein and gene-protein interactions in scientific literature, which will eventually facilitate the elucidation of the leukemogenic process.
While data from these sources will probably facilitate a complete and accurate

classification of AML and advance the elucidation of AML pathogenesis, it is to be kept in mind that due to the heterogeneity of AML, the number of samples concurrently analyzed is a critical factor and will remain a limiting step in this explorative process.

1. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531-7.
2. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347(25):1999-2009.
3. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530-6.
4. Haferlach T, Kohlmann A, Schnittger S, et al. Global approach to the diagnosis of leukemia using gene expression profiling. Blood 2005;106(4):1189-98.
5. Virtaneva K, Wright FA, Tanner SM, et al. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc Natl Acad Sci U S A 2001;98(3):1124-9.
6. Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30(1):41-7.
7. Schoch C, Kohlmann A, Schnittger S, et al. Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. Proc Natl Acad Sci U S A 2002;99(15):10008-13.
8. Debernardi S, Lillington DM, Chaplin T, et al. Genome-wide analysis of acute myeloid leukemia with normal karyotype reveals a unique pattern of homeobox gene expression distinct from those with translocation-mediated fusion events. Genes Chromosomes Cancer 2003;37(2):149-58.
9. Lacayo NJ, Meshinchi S, Kinnunen P, et al. Gene expression profiles at diagnosis in de novo childhood AML patients identify FLT3 mutations with good clinical outcomes. Blood 2004;104(9):2646-54.
10. Neben K, Schnittger S, Brors B, et al. Distinct gene expression patterns associated with FLT3- and NRAS-activating mutations in acute myeloid leukemia with normal karyotype. Oncogene 2005;24(9):1580-8.
11. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N Engl J Med 2004;350(16):1605-16.
12. Marcucci G, Radmacher MD, Ruppert AS, et al. Independent Validation of Prognostic Relevance of a Previously Reported Gene-Expression Signature in Acute Myeloid Leukemia (AML) with Normal Cytogenetics (NC): A Cancer and Leukemia Group B (CALGB) Study. In: ASH 2005. Atlanta: American Society of Hematology; 2005.
13. Heuser M, Wingen LU, Steinemann D, et al. Gene-expression profiles and their association with drug resistance in adult acute myeloid leukemia. Haematologica 2005;90(11):1484-92.
14. Callens C, Chevret S, Cayuela JM, et al. Prognostic implication of FLT3 and Ras gene mutations in patients with acute promyelocytic leukemia (APL): a retrospective study from the European APL Group. Leukemia 2005;19(7):1153-60.
15. Ross ME, Mahfouz R, Onciu M, et al. Gene expression profiling of pediatric acute myelogenous leukemia. Blood 2004;104(12):3679-87.
16. Alizadeh A, Eisen M, Davis RE, et al. The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. Cold Spring Harb Symp Quant Biol 1999;64:71-8.
17. Agendia BV. 2005. (Accessed at http://www.agendia.nl)
18. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 2004;20(3):323-31.
19. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185-93.
20. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002;18 Suppl 1:S96-104.

21. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31(4):e15.
22. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 2001;98(1):31-6.
23. Affymetrix. Guide to probe logarithmic intensity error (PLIER) estimation; 2005.
24. Cope MC, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 2004;20(3):323-31.
25. Van Putten W, Verhaak RGW, P.J.M. V, Van Wieringen W. Using correlations to reduce noise in gene expression analysis using Affymetrix GeneChip; 2005.
26. Beekman R, Touw IP. In; 2006.
27. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B. Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Res 2003;31(6):1753-64.
28. Aerts S, Van Loo P, Thijs G, et al. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Res 2005;33(Web Server issue):W393-6.
29. Haverty PM, Hansen U, Weng Z. Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. Nucleic Acids Res 2004;32(1):179-88.
30. Wouters B, Erpelinck CA, Valk PJ, Verhaak RGW, Lowenberg B, Delwel R. Unique gene expression profiles of AML patients with CEBPα mutations. In: ASH 2005. Atlanta: American Society of Hematology; 2005.
31. Segal E, Shapira M, Regev A, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003;34(2):166-76.
32. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34(3):267-73.
33. Rhodes DR, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 2004;101(25):9309-14.
34. Hwang D, Rust AG, Ramsey S, et al. A data integration methodology for systems biology. Proc Natl Acad Sci U S A 2005;102(48):17296-301.
35. Hwang D, Smith JJ, Leslie DM, et al. A data integration methodology for systems biology: experimental verification. Proc Natl Acad Sci U S A 2005;102(48):17302-7.
36. Calvano SE, Xiao W, Richards DR, et al. A network-based analysis of systemic inflammation in humans. Nature 2005;437(7061):1032-7.
37. Nikolsky Y, Nikolskaya T, Bugrim A. Biological networks and analysis of experimental data in drug discovery. Drug Discov Today 2005;10(9):653-62.
38. Jelier R, Jenster G, Dorssers LC, et al. Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics 2005;21(9):2049-58.
39. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. Nature 1993;364(6437):555-6.
40. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251(4995):767-73.
41. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci U S A 1994;91(11):5022-6.
42. Pinkel D, Segraves R, Sudar D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998;20(2):207-11.
43. Fitzgibbon J, Smith LL, Raghavan M, et al. Association between acquired uniparental disomy and homozygous gene mutation in acute myeloid leukemias. Cancer Res 2005;65(20):9152-4.
44. Raghavan M, Lillington DM, Skoulakis S, et al. Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. Cancer Res 2005;65(2):375-8.
45. Weinmann AS, Farnham PJ. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. Methods 2002;26(1):37-47.
46. Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding

proteins. Science 2000;290(5500):2306-9.
47. Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 2002;16(2):235-44.
48. Esteller M, Fraga MF, Paz MF, et al. Cancer epigenetics and methylation. Science 2002;297(5588):1807-8; discussion -8.
49. Huang TH, Perry MR, Laux DE. Methylation profiling of CpG islands in human breast cancer cells. Hum Mol Genet 1999;8(3):459-70.
50. Early P, Rogers J, Davis M, et al. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. Cell 1980;20(2):313-9.
51. Perry RP, Kelley DE. Immunoglobulin messenger RNAs in murine cell lines that have characteristics of immature B lymphocytes. Cell 1979;18(4):1333-9.
52. Rogers J, Early P, Carter C, et al. Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. Cell 1980;20(2):303-12.
53. Nitta E, Izutsu K, Yamaguchi Y, et al. Oligomerization of Evi-1 regulated by the PR domain contributes to recruitment of corepressor CtBP. Oncogene 2005;24(40):6165-73.
54. Affymetrix. GeneChip exon array system; 2005.
55. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 2004;5(7):522-31.
56. Martinelli G, Ottaviani E, Buonamici S, et al. Association of 3q21q26 syndrome with different RPN1/EVI1 fusion transcripts. Haematologica 2003;88(11):1221-8.
57. Fazi F, Rosa A, Fatica A, et al. A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. Cell 2005;123(5):819-31.
58. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422(6928):198-207.
59. Griffin TJ, Gygi SP, Ideker T, et al. Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. Mol Cell Proteomics 2002;1(4):323-33.
60. Garraway LA, Widlund HR, Rubin MA, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature 2005;436(7047):117-22.
61. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29(4):365-71.
62. Microarray Gene Expression Data Society. (Accessed at http://www.mged.org.)
63. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 2002;3(8):SOFTWARE0003.
64. Sherlock G, Hernandez-Boussard T, Kasarskis A, et al. The Stanford Microarray Database. Nucleic Acids Res 2001;29(1):152-5.
65. Saeed AI, Sharov V, White J, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 2003;34(2):374-8.
66. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863-8.
67. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 2003;19(18):2448-55.
68. The R Project for Statistical Computing. 1996. (Accessed at http://www.r-project.org.)
69. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.
70. Bullinger L, Valk PJ. Gene expression profiling in acute myeloid leukemia. J Clin Oncol 2005;23(26):6296-305.
71. Valk PJ, Delwel R, Lowenberg B. Gene expression profiling in acute myeloid leukemia. Curr Opin Hematol 2005;12(1):76-81.
72. Valk PJ, Vankan Y, Joosten M, et al. Retroviral insertions in Evi12, a novel common virus integration site upstream of Tra1/Grp94, frequently coincide with insertions in the gene encoding the peripheral cannabinoid receptor Cnr2. J Virol 1999;73(5):3595-602.
73. Verhaak RG, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in

acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. Blood 2005;106(12):3747-54.

Samenvatting

Hematopoïese, of de vorming van functionele bloedcellen, is een proces wat plaats vindt in het beenmerg. Hematopoïetische stamcellen ondergaan cycli van deling en differentiatie waarin de functionele eindcellen, zoals rode bloedcellen, bloedplaatjes en witte bloedcellen, worden gevormd. Leukemie is een ziekte waarbij de stamcellen abnormale processen van deling in combinatie met een stop van de differentiatie ondergaan, waardoor er de vorming van functionele eindcellen wordt belemmerd. In het geval van acute myeloïde leukemie (AML) is er een afwijking in de tak van bloedcelvorming waar onder andere rode bloedcellen, bloedplaatjes en granulocyten worden gevormd.

De ontsporing van hematopoïetische stamcellen met AML als gevolg wordt veroorzaakt door abnormaliteiten in het genoom, zoals chromosomale fusies, deleties en mutaties. De klinische prognose wordt momenteel bepaald aan de hand van de aan- of afwezigheid van (combinaties van) abnormaliteiten.

Het belangrijkste gevolg van genomische afwijkingen is de abnormale transcriptie van genen naar mRNA. Met behulpvan gen expressie profilering, door middel van microarrays, kunnen de transcriptie niveaus van duizenden genen simultaan worden bepaald. In hoofdstuk 2 is een onderzoek beschreven waarin met gen expressie profilering is toegepast op 285 beenmerg monsters van *de novo* AML patiënten, voor het bepalen van prognose. Verschillende bekende prognostische groepen, zoals t(8;21) en inv(16) konden worden geidentificeerd, alsmede een nieuwe prognostisch relevante groep van patiënten met een relatief slechte prognose (cluster 10).Hoofdstuk 2 laat zien dat gen expressie profilering in staat is om de huidige technieken voor het bepalen van prognose te vervangen, en prognose te verbeteren.

Om expressie niveaus te bepalen en te normalizeren met behulp van microrarrays worden verschillende statistische technieken toegepast. Deze kunnen leiden tot verschillende analyse uitkomsten. In hoofdstuk 3 wordt ingegaan op het effect van verschillende statistische technieken op de uitkomst van analyse. Methoden die een globale indruk van de data geven, zoals clustering en predictie analyse, worden niet erg beïnvloedt. Methoden die op de individuele data werken, zoals detectie van differentiële expressie, zijn daarentegen wel gevoelig voor de verschillende manieren van pre-processen. Door het ontbreken van een gouden standaard kan niet worden vastgesteld welke pre-processing methode het meest accuraat is.

Mutaties in het gen nucleophosmine (*NPM1*) zijn aanwezig in 35% van AML patiënten. Mutaties in *NPM1* correleren met hoge witte bloedcel aantallen, een normaal karyotype en mutaties in het gen *FLT3* (hoofdstuk 4). *NPM1* mutaties worden zelden gevonden in combinatie met mutaties in de genen *CEBPa* en *NRAS*, en zijn negatief gecorreleerd met een leeftijd onder de 35 jaar. Mutaties in *NPM1* zijn geassocieerd met een specifiek expressie profiel en multivariate analyse toont aan dat patiënten met mutaties in *NPM1* een relatief gunstige prognose hebben.

Retrovirale insertionele mutagenese is een krachtige techniek voor het identificeren van genen betrokken bij de ontwikkeling van kanker in muizen. De relevantie van de gevonden genen voor humane ziekten is echter onduidelijk. Hoofdstuk 5 laat zien dat genen geïdentificeerd met behulp van retrovirale insertionele mutagenese significant vaker differentieel tot expressie komen, dan overige genen. Genen die op grotere afstand van de plaats van de insertie liggen, tot 1 megabase, komen niet

significant vaker differentieel tot expressie. Onze data ondersteunen de validiteit van retrovirale insertionele mutagenese voor het detecteren van genen betrokken bij humane ziekten.
Gen expressie profielen wordt traditioneel weergegeven in sample versus gen heatmaps, waarin expressie niveaus worden weergegeven op een groen tot rode schaal. In een dergelijke weergave is er geen overzicht van gecorreleerde profielen en is het moeilijk om sample karakteristieken weer te geven. Hoofdstuk 6 beschrijft een methode voor de efficiënte en accurate visuele weergave van de resultaten van een clusteranalyse. In deze weergave worden correlaties gecombineerd met sample karakteristieken.

Conclusies

Genoom-brede gen expressie studies zijn veelbelovend voor het verbeteren het van prognose, en deze techniek zal in de toekomst de bestaande technieken grotendeels vervangen. Door het combineren van data uit gen expressie profilering studies met data van andere studies, zoals studies van mutaties of integraties in specifieke genen, danwel high-througput studies, zoals studies waarin SNP profielen worden geanalyzeerd, kunnen belangrijke nieuwe inzichten in de pathogenese van leukemie worden verkregen.

Abbreviations

AGM	Aorta-gonad-mesonephros
AML	Acute Myeloid Leukemia
AML1	Acute myeloid leukemia 1 gene
APL	Acute Promyelocytic Leukemia
BAALC	Brain and acute leukemia, cytoplasmic gene
bp	Base pair
CBF	Core binding factor
CBFB	Core-binding factor beta gene
CD34	CD34 antigen gene
cDNA	DNA complementary to RNA
CEBPA	CCAAT/enhancer binding gene
CGH	Comparative genomic hybridization
ChiP	Chromatin immuno-precipitation
DNA	Deoxyribonucleic Acid
EPO	Erythropoietin gene
EST	Expressed Sequence Tag
ETO	Eight twenty-one gene
EVI1	Ecotropic viral integration 1 site gene
FAB	French-American-British classification
FLT3	FMS-like tyrosine kinase 3 gene
GCRMA	GC robust multi-array average
GEP	Gene expression profiling
GCSF	Granulocyte colony stimulating factor gene
GMCSF	Granulocyte macrophage-colony stimulating factor gene
HSC	hematopoietic stem cells
IL-*n*	Interleukin *n* gene
Inv	Inversion
ITD	Internal tandem duplication
Kb	Kilobase
KL	Kit ligand gene
LOH	Loss of heterozygosity
MAC1	Macrophage antigen 1
MADEx	Microarray data explorer
MAGE-ML	Microarray and Gene Expression Markup Language
MAS5.0	Microarray Analysis Suite version 5.0
MCSF	Monocyte colony stimulating factor gene
MIAME	Minimum Information About Microarray Experiments
MITF	Microphthalmia-associated transcription factor
MLL	Mixed leukemia lineage gene
MM	MisMatch probe
MPO	Myeloperoxidase gene
mRNA	Messenger RNA
MYH11	Myosin heavy chain 11 gene
PAM	Prediction analysis of microarrays
PLIER	Probe logarithmic intensity error
PM	Perfect Match probe

PML	Promyelocytic leukemia gene
RARA	Retinoic acid receptor alfa gene
RMA	Robust multi-array average
RNA	Ribonucleic Acid
SAM	Significance analysis of microarrays
SCA1	Stem cell antigen 1 gene
SCF	Stem cell factor gene
SNP	Single nucleotide polymorphism
t	Translocation
TKD	Tyrosine kinase duplication
TP53	Tumor protein 53 gene
TPO	Thrombopoietin gene
WHO	World Health Organization
WT1	Wilms'tumor suppressor gene

Dankwoord

Heel veel mensen zijn op één of andere wijze betrokken geweest bij het tot stand komen van mijn proefschrift. Allereerst wil ik mijn promotor, Bob Löwenberg, bedanken voor de gelegenheid om te promoveren binnen zijn afdeling. Bob, daarnaast wil ik je ook graag bedanken voor je persoonlijke betrokkenheid en visie bij alle projecten.
Ik ben Peter Valk, co-promotor, veel dank verschuldigd voor zijn enorme bijdrage en steun. Peter, wat mij betreft had onze samenwerking niet beter kunnen verlopen. Ik hoop ook hierna kontakt te blijven houden zodat ik van je kan blijven leren. Bedankt voor alles.
Graag wil ik de leden van de kleine commissie, Prof.dr. Pieters, prof.dr. Van der Spek en dr. Van 't Veer, bedanken voor hun bijdrage en het beoordelen van het manuscript. Peter van der Spek wil ik graag nogmaals danken voor het in contact brengen met de afdeling Hematologie – zonder dat was dit dankwoord er überhaupt niet geweest.

Ruud Delwel wil ik graag danken voor zijn scherpe inzichten en persoonlijke gesprekken en het liefdevol opnemen in zijn werkbespreking en ook een beetje zijn groep. Van Ivo Touw heb ik veel geleerd over de wetenschap en bevlogenheid, waarvoor dank. En misschien moeten we het nog een keer over echte muziek hebben, danwel echt over muziek. 'Wervelwind' Marieke von Lindern weet dingen zo inspirerend over te brengen, daar moet je wel van houden. Succes met Rosetta en je geliefde ANOVA's.

I am blessed with two beautiful and fantastic paranymfs – 'parabitch' Meri Alberich Jorda and 'paradude' Eric van den Akker. With both (and also the three of us) we had brilliant times – I hope to have many more of them. Thank you for your support and for bearing with me in those 60 minutes. Meri, we are going to rock Boston hard! Dudemeister, yin&yang, autistische GVR, bedankt voor vele goede tijden en duizenden interessante feitjes over muziek en voetbal.

'Mijn' studenten, Mathijs Sanders en Maarten Bijl, ben ik ontzettend dankbaar voor al het prachtige werk dat ze verzet hebben. Thijs, veel succes in Delft, mocht je ooit weer een stageplek nodig hebben en een Amerikaans avontuur willen... Metal forever! Maarten, succes in de music-bizz, 50po r0x0rs! Rock on.

Stefan Erkeland, we hebben serieus gewerkt en veel gelachen tijdens onze samenwerking. En ja, het stuk is nog heel aardig terecht gekomen!! Bedankt voor de leuke tijd in Boston, jammer dat jij weggaat nu ook die kant op kom. En is dat wel helemaal toeval...? Succes met alles wat je nog gaat doen.

Dick de Ridder, samen hebben we een pittige klus geklaard. Je bent hierbij zeer motiverend geweest, en altijd duidelijk en geduldig met uitleg. Bedankt hiervoor.

Bart Aarts, bedankt voor de gesprekken over alles wat boeit in het leven – succes met alle autopsies en pulk-acties die gaan komen in je nieuwe baan. Walbert Bakker en Annemiek Broyl, dank voor de prima kamergenoten die jullie waren. Veel lol

gehad met jullie, en een goede buur is beter dan een verre vriend! Annemiek, succes met je verdere onderzoek. Alle (ex-)kamergenoten ook dank voor het (min of meer) tolereren van mijn muziek ;).

Fokke, bedankt voor de mooie Sparta sticker boven mijn bureau, misschien nog een keer vogelen? MADEx-groupie no.1 Bas, succes met GEPpen en bedankt voor het corrigeren van mijn SAM uitleg. Oud-collega's Sahar Khosrovani, Dominik Spensberger, Nazik Rayman, bedankt. Godfrey Grech, good luck back at Malta. Renee Beekman, succes alvast met promoveren straks.

Justine Peeters, thanks for your support during bad and good times. Good luck finishing your thesis, and I hope we will be in contact for a long time! Karlijn Schellekens, veel succes met alles wat je nog gaat doen. Ook alle andere mensen van de afdeling Bioinformatica, speciaal Mirjam van Vroonhoven en Bas Horsman: dank voor jullie samenwerking.

To all of my fellow PhD-students – especially Karishma Palande, Saman Abbas, Sanne Lugthart, Sophie Corthals, Andrzey 'Tony' Nieradka – good luck finishing your theses and with everything that you are going to do. Ana Guimaraes, prettig om samen met jou de laatste loodjes te doorstaan. Succes met je verdediging.

Dames van de diagnostiek – bedankt voor al jullie werk, zonder jullie data was dit nooit gelukt. Ook bedankt voor de steun bij mijn onhandige pogingen tot labwerk. Claudia Erpelinck en Antoinette Beijen, dank voor jullie fantastische werk met de chips. Aan alle mensen op het lab die niet genoemd staan, Delwel-groep, Touw groep, etc: bedankt voor de lol.

Jan van Kapel, over veel dingen zullen wij het nooit eens worden. Toch heb je me nooit teleurgesteld wanneer ik weer iets nodig had. Bedankt hiervoor. Ans Mannens, Jeanne Vlasveld, Eveline van Heese en Eveline Streefkerk: bedankt voor jullie belangrijke administratieve ondersteuning.

To all my friends outside the lab, in Nijmegen, Rotterdam and Boston, from BGW, Deoluzion, 'international people', Wacken, football, other things – I owe a lot to many of you. Without you this thesis would never have finished. I hope to see all of you in Boston, and you will definitely see me when I come back. Thankyoutha nkyouthankyou.

Waar moet je beginnen als je je ouders wilt bedanken? Jullie steun is van onschatbare waarde. Bedankt. Ook zus, broertjes en aanhangsels: bedankt voor inspiratie en motivatie. Opa en oma V. natuurlijk bedankt voor de ultieme inspiratie voor een academische carrière.

A metal heart is hard to tear apart. Party on!
R.

Curriculum Vitae

Roeland George Willehad Verhaak was born in Wijchen, the Netherlands, on September 29 1976. After finishing his VWO education at the Kottenpark College in Enschede in 1996, he started a curriculum Biomedical Health Sciences at the Catholic University Nijmegen (KUN, currently Radboud University). As part of this education, he followed majors in pathobiology and toxicology, and a minor in computer science. A toxicology internship, titled 'Mitochondrial toxicity of nuclease reverse transcriptase inhibitors, was completed at the Department of Pharmacology and Toxicology of the KUN under supervision of Dr. Roos Masereeuw. A second intership project, 'Development of a diagnostic marker of multiple sclerosis', was completed at the Department of Biochemistry, under supervision of Dr. Rinie van Boekel en Prof.dr. W. Van Venrooij. He obtained his Masters–degree in August 2000. After having started a project at the Department of Medical Informatics of the KUN in October 2000 in which he worked on structuring of temporal data, he switched to the bioinformatics company Dalicon BV in April 2002. At Dalicon, he worked as software engineer, with a particular focus at the database system SRS. In April 2003 he started a PhD-project at the Department of Hematology at the Erasmus MC in the lab of Prof.dr. Bob Löwenberg, supervised by Dr. Peter Valk. This work has been described in this thesis. From March 2006 until June 2006, he was a visiting scientist of the Department of Biostatistics and Computational Biology of the Dana-Farber Cancer Institute in Boston, supervised by Prof.dr. John Quackenbush. The author wil continue his academic career at the Broad Institute in Boston, a research collaboration of MIT, Harvard and its affiliated hospitals, and the Whitehead Institute.

List of publications

1. Valk PJ, **Verhaak RG**, Beijen MA, Erpelinck CA, Barjesteh van Waalwijk van Doorn-Khosrovani S, Boer JM, Beverloo HB, Moorhouse MJ, van der Spek PJ, Löwenberg B, Delwel R. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med. 2004 Apr 15;350(16):1617-28.

2. **Verhaak RG**, Goudswaard CS, van Putten W, Bijl MA, Sanders MA, Hugens W, Uitterlinden AG, Erpelinck CA, Delwel R, Löwenberg B, Valk PJ. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. Blood. 2005 Dec 1;106(12):3747-54.

3. Erkeland SJ*, **Verhaak RG***, Valk PJ, Delwel R, Löwenberg B, Touw IP. Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. Cancer Res. 2006 Jan 15;66(2):622-6.
These authors contributed equally

4. **Verhaak RG**, Staal FJ, Valk PJ, Löwenberg B, Reinders MJ, de Ridder D. The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies. BMC Bioinformatics. 2006 Mar 2;7:105.

5. **Verhaak RG**, Sanders MA, Bijl MA, Delwel R, Horsman S, Moorhouse M, Van der Spek PJ, Löwenberg B, Valk PJ. HeatMapper: Powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics. BMC Bioinformatics. 2006 Jul 12;7(1):337.

6. Steidl U, Rosenbauer F, **Verhaak RG**, Gu X, Gu X, Ebralidze A, Otu HH, Klippel S, Steidl C, Bruns I, Costa DB, Wagner K, Avaido M, Kobbe G, Valk PJ, Passegué E, Libermann TA, Delwel R, Tenen DG. Essential role of Jun family transcription factors in PU.1-induced leukemic stem cells. Nat Genetics. 2006. *Accepted for publication*.

7. Steidl U, Steidl C, Ebralidze A, Han HJ, Rosenbauer F, Koschmieder S, Wagner K, Kobayashi S, Schulz T, Becker A, O'Brien KB, Krauter J, Haase D, **Verhaak RG**, Delwel R, Truemper L, Kohwi-Shigematsu T, Griesinger F, Tenen DG. A distal single nucleotide polymorphism alters long-range regulation of the PU.1 gene by special AT-rich sequence binding protein 1 (SATB1) in acute myeloid leukemia. *Submitted.*

8. Bakker WJ, Van Dijk TB, Parren-van Amelsvoort M, Kolbus A, Steinlein P, **Verhaak RG**, Mak TW, Beug H, Löwenberg B, Von Lindern M. Differential regulation of Foxo3a target genes in erythropoiesis. *Submitted.*