Class Discovery in Acute Lymphoblastic Leukemia using Gene Expression Analysis

Uri Einav

M.Sc. Thesis submitted to the Scientific Council of the Weizmann Institute of Science

Research conducted under the supervision of
Prof. Eytan Domany

December 2003

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Eytan Domany, for his continuous fascinating guidance, advice and support throughout the course of this research.

I also wish to thank Professor Gideon Rechavi and Dr. Ninette Amariglio for their enlightening discussions.

I am grateful to the members of 'Domany's group' and especially to Omer Barad, Gaddy Getz and Yuval Tabach for fruitful discussions.

My deep gratitude goes to my family and particularly to Yael for their encouragement.

Contents

Abstract

Leukemia is not a single homogenous disease and can be divided into subtypes, based on the associated chromosomal abnormalities. Recently, using microarray data analysis, several research groups characterized the gene expression profiles of different, previously known, leukemia subtypes. Our aim was **class discovery**: to identify new partitions of leukemia samples, into novel sub-groups with no previously known common label, on the basis of their gene expression profiles. Using a combination of supervised statistical methods and unsupervised clustering, we analyzed 12 publicly available datasets of leukemia and other cancers. Surprisingly, we observed two unanticipated partitions of the leukemia patients to novel subgroups. These two partitions are induced by two different clusters of genes. In both partitions clear and sharp separation of the patients was demonstrated on the basis of the expression levels of the corresponding gene clusters. While for one cluster the biological meaning of this separation is still vague, we suggest an interesting interpretation to the partition induced by the expression levels of the genes of the other cluster: this interpretation strengthens previous epidemiological and molecular studies, which suggested a role for viral infection in the etiology of leukemia and other cancers.

1 Introduction

1.1 Leukemia

Leukemia is a type of cancer that arises in the blood and bone marrow. Bone marrow, a soft type of tissue that is found in the center of some bones, is the body's "blood factory." It produces immature cells that eventually develop into new blood cells.

1.1.1 Hematopoietic Differentiation

Blood cell development begins in the marrow with the formation of hematopoietic stem cells. These primitive cells are capable of developing into any kind of blood cell. The first step in this evolution, or differentiation, is into one of two slightly more mature types of stem cells: **lymphocytic** progenitor cells and **myeloid** progenitor cells (Figure 1). These cells then undergo further specialization. Lymphocytic stem cells mature into either T cells, B cells, or natural killer cells. Myeloid stem cells mature into erythrocytes (red blood cells); platelets (which clot the blood); monocytes (a type of white blood cell); or granulocytes (a group of white blood cells that includes neutrophils, basophils, and eosinophils). Each of these types of cell has a specific role in the functioning of the body.

1.1.2 Leukemia: unregulated proliferation of hematopoietic cells

A malignant transformation can happen at any stage of blood cell development. These abnormal blood cells are characterized by abnormal unregulated proliferation of one or more cells of the hematopoietic lineage. This abnormal behavior is due to the generation of somatic mutations which confer the affected cells a proliferative or survival advantage over the normal cells. The mutations include translocations, deletions and insertions that usually affect oncogenes or tumor suppressors[1].

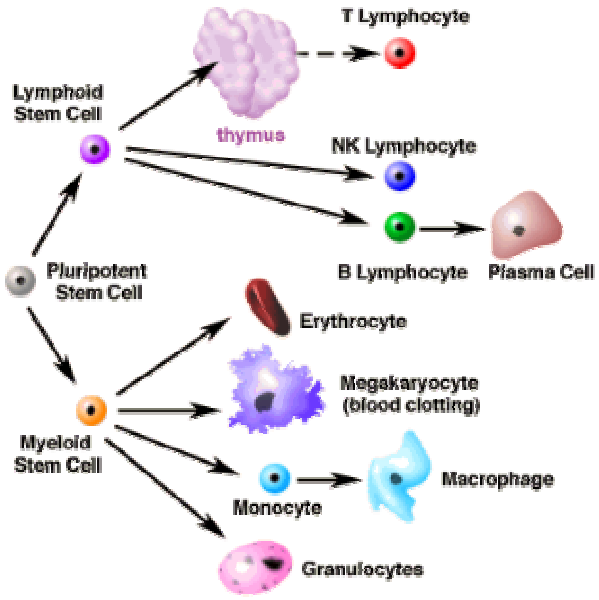


Figure 1 | Stem cell differentiation. Stem cells differentiate into the major players in the immune system (granulocytes, monocytes, and lymphocytes). Stem cells also differentiate into cells in the blood that are not involved in immune function, such as erythrocytes (red blood cells) and megakaryocytes ,for blood clotting (Adopted from http://www.biology.arizona.edu) .

1.2 Subtypes of leukemia

Most leukemias fall into one of two general groups: **myeloid leukemia** (about 60% of the cases) and **lymphocytic leukemia** (about 40%). People also classify leukemias according to whether they are **acute** (55%) or **chronic** (45%). In acute leukemias, the malignant cells, or blasts, are immature cells that are incapable of performing their immune system functions. The onset of acute leukemias is rapid (weeks), and, in most cases, fatal unless the disease is treated quickly. Chronic leukemias develop in more mature cells, which can perform some of their duties but not well. These abnormal cells also increase at a slower rate (years). Hence, there are four main types of leukemia: Acute Lymphocytic Leukemia (**ALL**), Chronic Lymphocytic Leukemia (**CLL**), Acute Myelogenous Leukemia (**AML**), Chronic Myelogenous Leukemia (**CML**). Each of these leukemia types consist of several subtypes. We will focus on the ALL.

1.2.1 Subtypes of ALL

There are several ways to classify ALL patients[2]. One method sub-divided the ALL into subtypes on the basis of morphological criteria into FAB (French-American-British) subtypes L1, L2, and L3. ALL cells are also classified by

immunophenotype, the lymphocytic cell line from which they are derived; in ALL disease arises in either B or T cells, which each have characteristic markers on their cell surfaces. Another important classification is based on the chromosomal abnormalities in the diseased cells.

1.2.1.1 Chromosomal abnormalities in ALL

The chromosomal changes, which contribute to the leukemiagenesis, contain: gain of chromosomes, loss of chromosomes, insertions, deletions and most prominent, chromosomal translocations (chromosome breaks and fusions). These chromosomal abnormalities occur in specific regions, and certain abnormalities are related to particular ALL subtypes. Chromosome translocations involve illegitimate recombination of normally separate genes[1] (Figure 2). This can result in dysregulation of expression of an oncogene by association with a powerful and constitutively active regulatory element[1,3]. More than 200 genes have been found to be involved in translocations in childhood leukemia[3], but many of these are rare and certain genes predominate (Table 1). The DNA breaks always occur in non-coding regions (introns) of genes. Breaks always occur, more or less randomly, within a limited region of these genes, but each patient's leukaemic cells have a unique (or clone specific) breakpoint in the DNA sequence, providing a specific, sensitive, and stable marker for tracking leukemic clones [4].

Table 1| Main biological subtypes and chromosome changes in childhood ALL (adopted from [3])

Subtype	Cell type involved	Chromosome abnormality	Molecular lesion
Acute lymphoblastic leukaemia (ALL)	B-cell progenitor–monocytic* (infants)	11q23 translocations	*MLL–AF4*, *MLL–ENL* and other fusions
	B-cell precursor	Hyperdiploidy	Increased gene dosage
		t(12;21)(p13;q22)	*TEL–AML1* fusion
		t(1;19)(q23;p13)	*E2A–PBX1* fusion
		t(9;22)(q34;q11)	*BCR–ABL* fusion
	T-cell precursor	1q deletion; t(1;14)(p32;q11)	*SIL–SCL* fusion

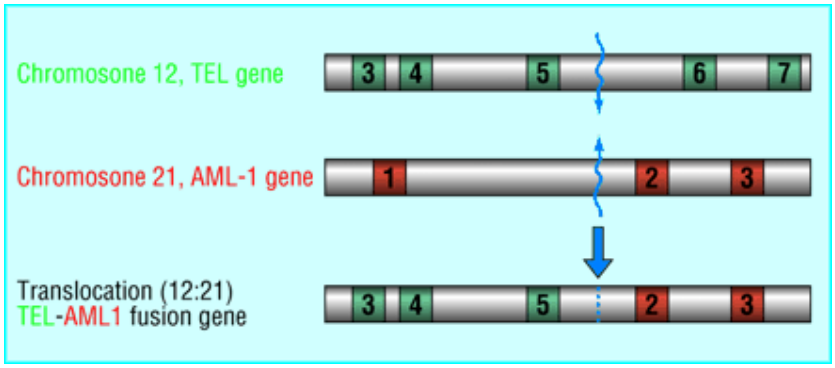


Figure 2 | Chromosomal translocation to form the TEL-AML1 fusion gene in childhood acute lymphoblastic leukaemia. The TEL and AML1 genes lie at the breaks and are brought together by the exchange. The genes break in non-coding (grey) regions between the coding regions (numbered, green or red), and re-joining of the two broken genes forms a novel fusion gene. (adopted from [5])

1.2.2 The frequency of leukemia subtypes as a function of age

There are marked differences in the frequency of specific subtypes as a function of age. Although the total number of hematopietic neoplasms in adults far exceeds the number seen in children, in the pediatric and adolescent populations these malignancies comprise almost 50% of all cancers, whereas in adults they comprise only 5%-8%[6] . In addition, the rate of certain subtypes of leukemia varies significantly between pediatric and adult patients. Acute lymphoblastic leukemia is the most common cancer seen in the pediatric population and accounts for greater than 50% of hematopoietic malignancies in this age group[6]. By contrast, ALL is a relatively rare leukemia subtype in adults, accounting for only 2%–3% of hematopoietic malignancies. Moreover, the rate of the secondary types of leukemia subtypes is also varying as a function of age. In ALL patients, MLL-AF4 gene fusion is much more frequent in infants than in children, and BCR-ABL gene fusion is more common in adults than in infants and children[5] (Figure 3).

One of the main datasets we analyzed in this study was derived from leukemic blasts of childhood ALL[7].

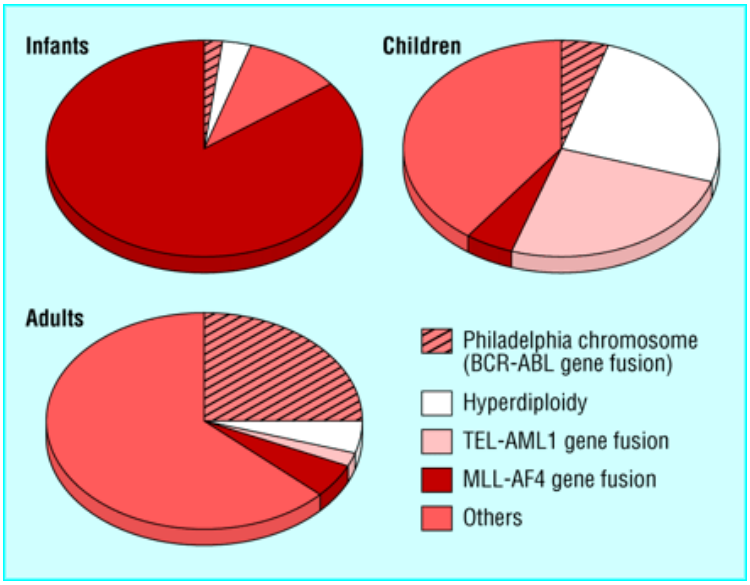


Figure 3 | Major molecular subtypes of acute lymphoblastic leukemia: in infants (<1 years old) children (2-10 years old) and adults. (adopted from [5])

1.3 Etiology of childhood leukemia

As mentioned above, childhood leukemia is not one homogenous disease and can be divided into subtypes, based on the chromosomal abnormalities. These include chromosome translocations, resulting in the generation of chimeric or fusion genes[1] and change in chromosome number (hyperdiploidy or hypodiploidy). Genes involved in these abnormalities have very diverse functions, but it seems that most of them are involved in some critical stage of cell growth, development, or survival[1].

There is now evidence[8,9] that the chromosome translocations are often the first of initiating events of leukemia, occurring **prenatally** during fetal development. This evidence comes mainly from studies of identical twin infants with TEL-AML1 translocation[8,9]. Analysis of pairs of identical twins with concordant acute lymphoblastic leukaemia shows that leukemic cells from both twins share the identical breakpoints in TEL and AML1 genes[8,9]. This contradicts the statement made above, that each patient's leukemic cells have a unique

breakpoint in the DNA sequence. Since gene breakpoints in leukemic cells are not inherited, the only plausible explanation for twin leukemias sharing the same gene breakpoints is that the chromosomal breaks generating the fusion gene must have occurred just once, in one blood stem cell, in one twin in utero. But, is this single chromosome translocation itself enough to generate leukemia?

It turns out that for identical twins aged 2-6 years, with acute lymphoblastic leukemia, the concordance rate of leukaemia is considerably low, around 5% only[10]; meaning that in 95% of the pairs, who shared the very same translocation, one twin did not suffer from leukemia. This indicates the **need for some additional postnatal event(s)**. For this additional event(s) there is a 1 in 20 chance (although it still represents a 100-fold extra risk of leukaemia for the twin of a patient with ALL).

1.3.1 The "two hit" model

The above finding suggests, at a minimum, a "two hit" model for the etiology of childhood leukemia. If this model of leukemia development is correct, then, for every child with acute lymphoblastic leukaemia diagnosed, there should be at least 20 healthy children who have had a chromosome translocation. This possibility has been investigated by screening unselected samples of newborn cord blood for fusion genes. About 600 samples have been screened, and around 1% have a leukaemic TEL-AML1 fusion gene[11]. This 1% represents 100 times the cumulative rate or risk of acute lymphoblastic leukaemia (with a TEL-AML1 gene). It turns out that the real bottleneck in development of acute lymphoblastic leukaemia therefore seems to be a strict requirement for a second "hit" after birth – in other words, additional chromosomal or molecular abnormalities are therefore needed for childhood leukaemia to develop.

1.3.2 The viral infection theory

Epidemiological evidence suggests that ionizing radiation and certain chemicals (such as benzene) may play a part in the development of some subtypes of leukaemia and lymphoma in adults and children[12]. In addition to these agents, it was suspected that common **childhood infections** contribute to the etiology

of childhood leukemia, in particular ALL. Most spontaneous leukemias in domesticated animals are related to viruses[12]. Several human lymphoid malignancies are associated with infectious agents: Burkitt's lymphoma with Epstein-Barr virus[13], ATL with HTLV-I[14], body-cavity lymphoma with human herpes 8[15], B-cell non-Hodgkin Lymphoma with hepatitis C[15] and gastric MALT lymphoma with Helicobacter Pylori[16].

The infectious etiology hypothesis is based on two distinct but complementary schools of two British researchers: **M. Greaves** and **L. Kinlen**. Both claim that infection in previously unexposed individuals may cause dysregulated immune response. This response, added upon the innate chromosomal abnormality, is blamed to be the "second hit" which contributes to leukemogenesis. They differ mainly in the suggested reason of the delayed exposure. The Kinlen theory[17], based on transiently increased rates of leukemia in geographical clusters, suggests that population mobility and mixing result in infection occurring in susceptible, previously unexposed individuals. Several epidemiological studies supported the population mixing theory[18,19]. The alternative "delayed infection" hypothesis[20-22] focuses on the timing of common childhood infections and claims that leukemia are associated with a lack of exposure in infancy.

Many common or endemic infections are encountered around birth through the mother, or during infancy through breast milk or other siblings or contacts. Such early exposures to the infectious agent should occur in the context of the immune protection that derives from the mother's transplacental antibodies and from breast milk thereafter. Early exposures modulate the infant's immunological repertoire and prepare it for future potential exposure. The changes in lifestyles in developed countries, in particular in high socioeconomic levels, including child-rearing and breastfeeding practices, compromise this evolutionary adaptation of the immune system. Pregnant women may not have been exposed to some infections and therefore will not provide immune protection to the infant. Moreover, the withdrawal of prolonged breastfeeding compromises immune protection and eliminates early oral transmission of infectious agents[23] .

Lack of early exposure leaves the immune system unmodulated, and later infection with common infectious agents might lead to inappropriate immune response. It was suggested that in some children abnormal immunological response to common infection may increase the risk of leukemia and it is postulated that it is the aberrant response to infection that promotes the crucial second, postnatal event[22].

Dysregulated immune response upon delayed exposure to infection is blamed to contribute to leukemogenesis. The dysregulated response to infection is suggested to provide, probably indirectly, proliferative or apoptotic stress to the bone marrow, leading to the complementary essential "hit"[22]. The exposure is predicted to occur proximally to clinical disease[5], suggesting that a "smoking gun" can be identified when leukemic cell samples are studied.

Despite intense research[24,25], no direct biologic supportive evidence, such as identification of microbial agent sequences, was provided. Similarly, no epidemiologic data linking any specific pathogen to ALL development were described. Several anecdotal reports described rare cases of ALL diagnosis preceded by a preleukemic phase known as pre-ALL in association with EBV or parvo B19 infection[26,27].

1.4 The importance of class discovery

Pediatric acute lymphoblastic leukemia is one of the great success stories of modern cancer therapy, with contemporary treatment protocols achieving overall long-term event-free survival rates approaching 80%[28]. This success has been achieved, in part, by using risk-adapted therapy that involves tailoring the intensity of treatment to each patient's risk of relapse. This approach was developed following the realization that pediatric ALL is a heterogeneous disease consisting of various leukemia subtypes that differ markedly in their response to chemotherapy[29]. Critical to the success of this approach has been

the accurate assignment of individual patients to specific leukemia subtypes. The underlying genetic abnormalities in these leukemia subtypes influence the response to cytotoxic drugs. Moreover, understanding of the molecular biology of certain subtypes can lead to smarter drug design, improved clinical efficacy and potentially more acceptable side-effects. While traditional cytotoxic cancer treatments such as chemotherapy or radiation therapy kill all dividing cells, these drugs can act on a molecular target by a mechanism that is more specific to particular leukemia subtype. Such a drug, *Gleevec*, was approved by the FDA in 2001 for the treatment of the BCR-ABL subtype[30]

Another implication of class discovery may be the recognition of a class of patients, who all share a certain causative agent in the pathogenesis of leukemia. This causative agent may contribute to the development of leukemia (in addition to the inborn chromosomal abnormality). The importance of the finding of such an agent on therapy and prevention of childhood leukemia is enormous. As mentioned, several human malignancies are associated with infectious agents. Some of them may be prevented or treated by vaccines designed to induce appropriate immune responses. Several efforts are currently carried out, in order to develop such a vaccine; among them: HPV vaccines for HPV-associated head and neck squamous cell carcinoma[31,32] and EBV vaccines for Hodgkin's disease (lymphoma)[33].

Naturally, the usage of such drugs and vaccinations is based on the accurate assignment of individual patients to a particular known and novel leukemia subtype. One hopes that the characteristics of each subtype can be determined using microarray studies.

1.4.1 A molecular-targeted drug: BCR-ABL and Gleevec

Traditional cytotoxic cancer agents have serious side effects such as nausea, weight loss, hair loss and severe fatigue that result from their lack of specificity in killing cells. Gleevec was designed as an inhibitor of a specific receptor associated with BCR-ABL, and so produces less severe side effects than other cancer agents[30]. The BCR-ABL translocation occurs at the site in the genome of a protein tyrosine kinase called ABL, creating the abnormal BCR-ABL protein,

a fusion of the ABL gene with another gene called BCR. The kinase activity of ABL in the BCR-ABL fusion is activated and unregulated, driving the uncontrolled cell growth. White blood cells containing the BCR-ABL mutation become able to proliferate in the absence of growth factors they normally require. Gleevec inhibits ABL kinase activity (Figure 4), helping to reverse uncontrolled cell growth. The activation of BCR-ABL also represses apoptosis through induction of anti-apoptosis factors such as Bad, allowing transformed cells to divide.

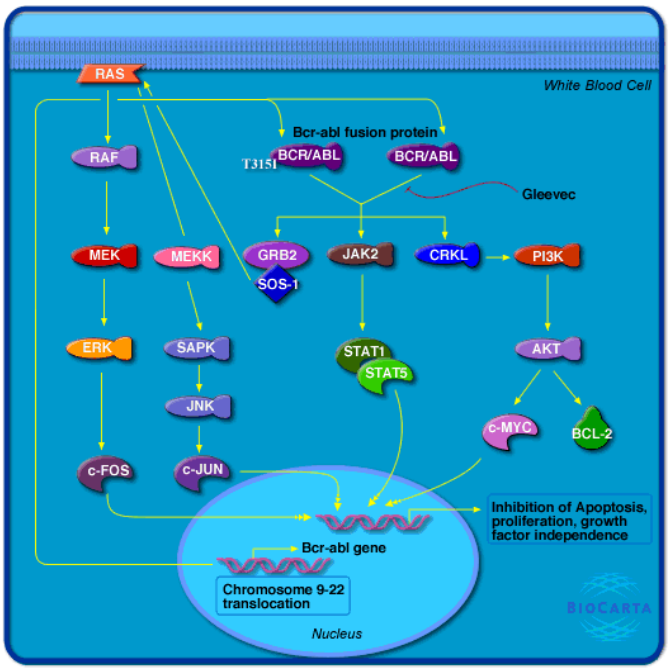


Figure 4| Inhibition of cellular proliferation by Gleevec in BRC-ABL subtype. (adopted from www.biocarta.com)

1.5 Recent leukemia microarray studies

Recently, using microarray data analysis, several research groups characterized the gene expression profiles of different known leukaemia subtypes and even revealed new ones[7,34-36]. Golub et al.[34] analyzed the gene expression of 38 acute leukemia samples and found groups of genes that distinguish ALL from acute myeloid leukemia (AML). Armstrong et al.[35]

analyzed 72 samples and showed that ALL with MLL translocations have a distinct gene expression profile and can be distinguished as a unique subtype of leukemia (MLL), separable from both ALL with no MLL translocation, and AML. Rozovskaia et al.[36] examined the gene expression of 60 ALL and AML. They found that 2/3 of the genes that separate MLL patients with t(4,11) translocations from the other ALL are, in fact, sensitive to the differentiation stage (mostly early for MLL). Only one third of the separating genes respond to the translocation. Yeoh et al.[7] analyzed the gene expression of 360 pediatric ALL patients and demonstrated that using the microarray platform, patients can be accurately classified to the relevant ALL subtypes (*T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL and hyperdiploid >50* chromosomes). In addition, a novel ALL subgroup was identified based on its unique expression profile. A different attempt was preformed trying to predict relapse, and indeed within some of these subgroups, distinct expression profiles that can predict relapse were identified.

1.6 Class Discovery – Methodologies

1.6.1 Laboratory techniques

Class discovery refers to defining previously unrecognized tumor subtypes[34]. Historically, class discovery was a difficult task, because it has relied on specific biological insights. The traditional techniques were prolonged procedures, which typically evolved through years of hypothesis-driven research. A more recent technique, FISH – Florescence-In-Situ-Hybridization uses fluorescent DNA probes that hybridize to specific portions of chromosomes. These probes can be used to identify certain parts of chromosomes, and to count the number of copies of each chromosome within tumor cells (Figure 5). A more effective technique, known as Spectral KaryotYping (SKY) is a technology based on FISH that is combined with another technology called spectral analysis. SKY, like FISH, employs florescent DNA probes that attach themselves to parts of chromosomes. The tagged portion of each chromosome appears in a different color, creating a multi-color

pattern that distinguishes one chromosome from another (Figure 5) [1,37]. Using these techniques, one can reveal previously unknown genetic abnormalities (as hyperdiploidity or novel translocations).

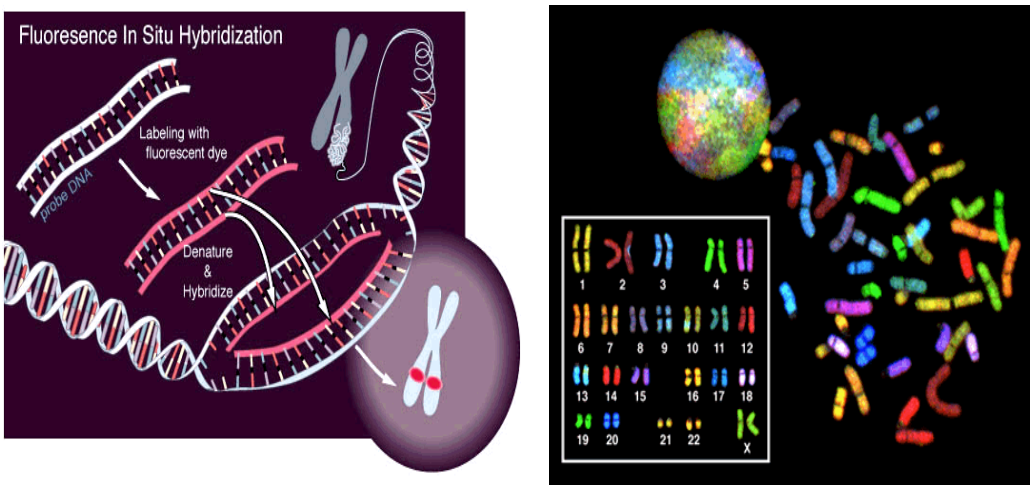


Figure 5| **Left**: The molecular basis of the FISH (Florescence In Situ Hybridization) technique. **Right:** The multi-color pattern of Sky - Spectral Karyotyping (adopted from http://www.accessexcellence.org).

1.6.2 Class discovery by gene expression analysis

An alternative approach may be based on global gene expression analysis, instead of applying labor intensive molecular examinations. As mentioned above, using microarray data analysis, several research groups have recently identified some known and even some previously unrecognized leukemia subtypes, based on common gene expression patterns [7,34-36]; each of the leukemia subtypes were identified by distinct gene expression profiles. These profiles may also serve as new and low-cost diagnostic tools. In addition, a detailed examination of these profiles may suggest candidate genes and proteins against which novel therapeutic drugs may be developed.

The class discovery problem involves two issues: first, using an unsupervised method such as clustering to group leukemia samples by their gene expression levels, and second, the determination whether putative classes produced by such a clustering algorithm are biologically meaningful.

In order to deal with both issues, we used a method which combined a clustering algorithm and other statistical tests and involved the analyses of several different datasets which were produced by several different research groups[34-36,38-45]; among them are 4 datasets of leukemia patients and 8 datasets of other types of cancer [38-45]. We employed the following multi-steps iterative approach, combining unsupervised and supervised techniques, consisting of several consecutive steps (Figure 6):

Step 1 – Discover a novel subgroup of samples within a known subtype

The aim of this step is to identify a new partition of the samples, into previously unrecognized sub-groups, on the basis of the gene expression profiles. The input for this step is gene expression data of leukemic patients. We start with one leukemic subtype only so that no 'inter-subtype' noise would affect the results. In order to perform a totally 'blind' analysis, an unsupervised technique is required. For this end we use CTWC, a bi-clustering method (which was developed in our lab, see Methods section). The result of this step is (hopefully) a group of co-expressed genes whose expression levels separate the studied subtype of leukemia into two unrecognized sub-groups. If indeed was found, such a group of genes is thought to participate in a particular biological pathway.

Step 2 – Refine the gene list

The two unfamiliar sub-groups of patients are identified by distinct expression levels of a certain set of co-expressed genes. However, there are, probably, additional genes that are also likely to be involved in the same pathway. In the same manner, some of the genes within the cluster may have been mistakenly included. Therefore, a refinement and an expansion of the cluster of genes are required. This may be done by standard supervised tests; the two novel sub-groups are treated as "ground truth" and taken as the basis for the test. The resulting refined list of genes is expected to differentiate between the two new sub-groups. The TNOM test is applied for this purpose (see Methods section).

Step 3 – Expanding the novel subgroup of samples

The novel sub-group we have at this point consists solely of one known subtype of leukemia. As a result, we might miss interesting insights; revealing the distribution of all given leukemia subtypes within the novel subgroup can shed biological light on the demonstrated phenomenon. Hence, we turn to expand the novel subgroup by analyzing other leukemia subtypes as well. For this aim, we use again a clustering algorithm (SPC), clustering samples from all subtypes using the refined list of genes. The output of this step is an expanded list of samples, which co-express these genes.

Step 4 – Refine the list of genes

To complete the refinement process, supervised analysis (TNOM) is performed again. This time, we try to find genes that separate the expanded list of samples (taken from all given subtypes of leukemia) from the remaining samples, on the basis of their expression values. At this point we expect to keep fixed the final refined list of genes that are expressed distinctively by the members of the novel subgroup.

At this point, we should pose hypotheses regarding the biological meaning of the novel subgroup of genes that generated the novel partition of the patients. We do so, on the basis of two types of information: the annotations of the genes and clinical data of the samples.

Step 5 – Other leukemia datasets

To substantiate the findings obtained from previous steps, we may test whether we can find a similar sub-division in other leukemia data sets[34-36], using the same set of genes. Similar sub-division may strengthen the significance of the results; on the other hand, difficulties in the ' transferability' of our results to other leukemia datasets can imply that our findings are an artifact specific only to the particular dataset studied. This step allows us to try to verify the hypothesis we posed in the previous step.

Step 6 – Other cancer datasets

Successful 'transfer' of our findings out of the leukemia limits can demonstrate a wider phenomenon, which is not necessarily limited to leukemia. For this reasons, we analyze, in addition to the leukemia datasets, several other datasets taken from lymphoma[40], prostate[41,44] various mixed tumors [39,42], ovary[43] , lung[38] and breast[45], always trying to induce an unknown separation, using the same novel cluster of genes . Individual analyses of certain datasets allow examination of specific aspects of the biological hypothesis (e.g. analysis of lymphoma datasets in order to examine whether the demonstrated phenomenon exists in similar types of tissues too).

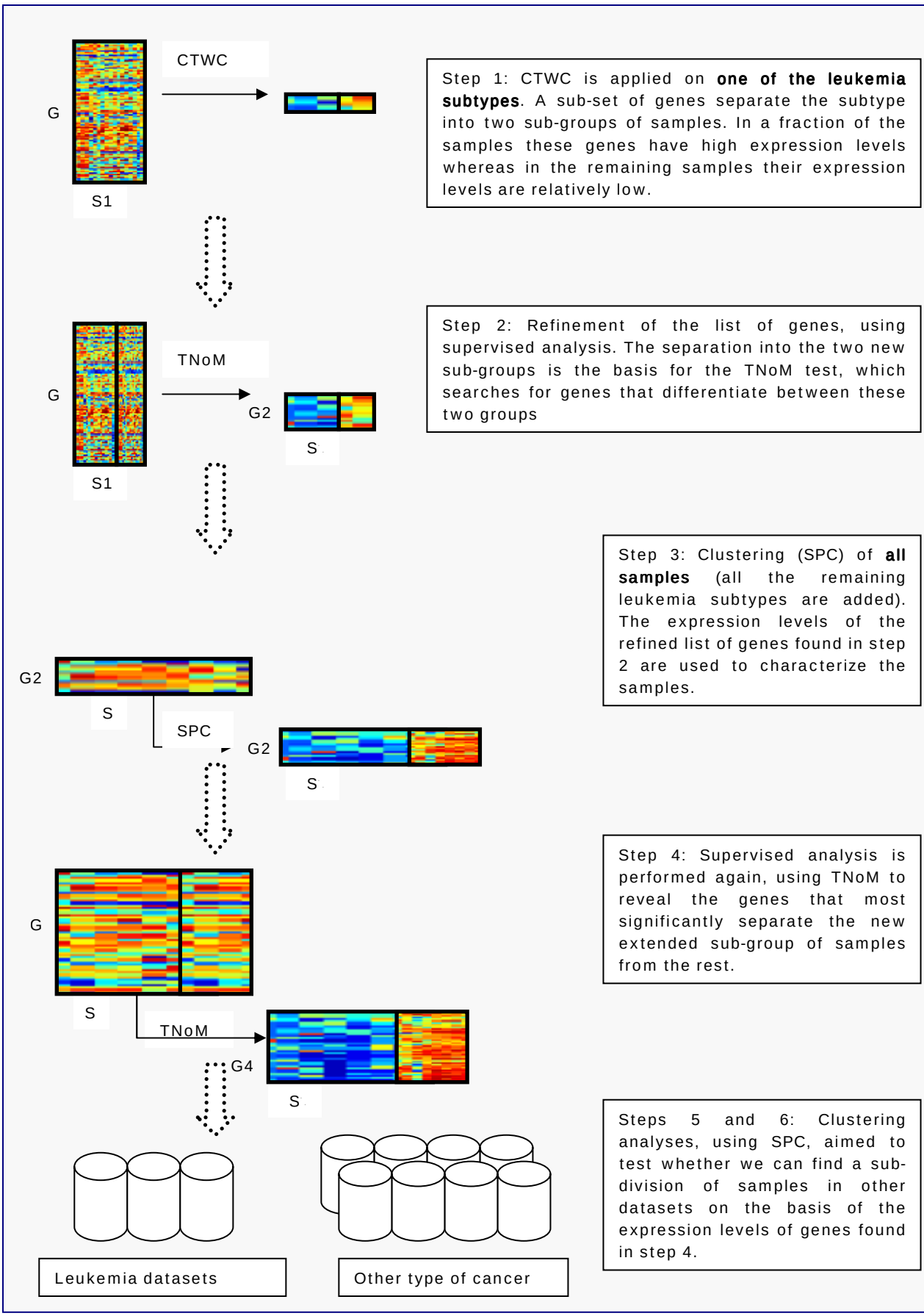


Figure 6 | The multi-step iterative approach we applied in this study.

Our procedure is reminiscent in spirit of the signature method[46]. This algorithm uses a set of genes as input (Figure 7); usually, genes that are known to be co-expressed. The algorithm identifies the samples in which the input genes are highly expressed. The score of each sample is the average change in the expression of the input genes. Only samples with a large absolute score are selected. Next, the algorithm searches the whole set of genes and selects those that show a significant and consistent increase of expression in the samples selected in the first stage. By doing so, the algorithm refines and extends the original set of genes.

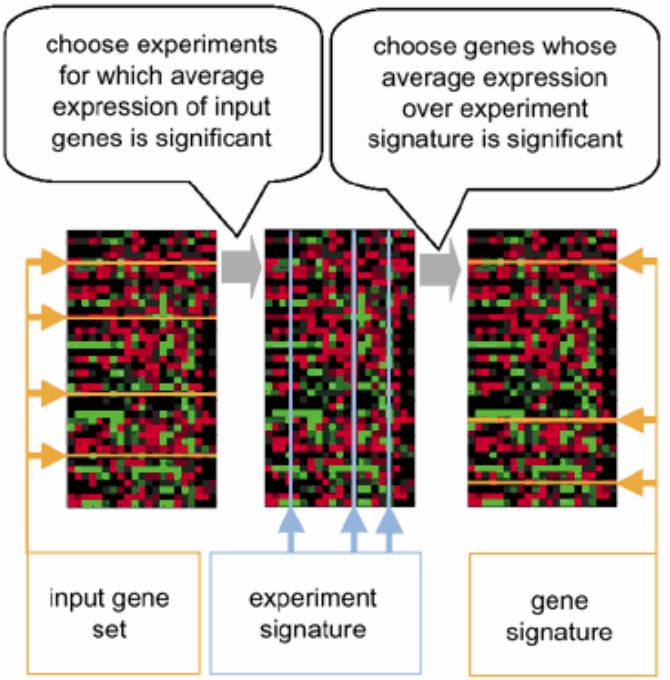


Figure 7| The signature Algorithm (adopted from [46]).

The signature method starts with a known/suspected set of genes, while our procedure searches for them in an unsupervised way. Moreover, the signature method analyzes all samples in the first step (analysis of each sample separately). We use a clustering algorithm, which takes into consideration the information from all given samples together, and therefore we believe that 'inter-subtype' noise can increase the Signal to Noise Ratio. We prefer to first search for a clue to the pathway within one subtype and then to expand it to the rest. Another crucial step we take is the 'transferability' test to other

datasets. This step can demonstrate the significance of the findings and to shed a light on the biological meaning which may remain obscure.

Applying our procedure on any given gene expression data allows us to reveal hidden subtypes in the data. Understanding the biological meaning of these subtypes is much more difficult, and depends also on the existence of relevant solid genomic, molecular or biological knowledge.

2 Results and discussion

Our aim was ***class discovery***: to identify new partitions of the samples, into sub-groups with no previously known common label, on the basis of the expression profiles of a group of genes with correlated expression levels. To this end we applied the CTWC method (see methods section) on the data of *Yeoh et al*[7]. The expression levels of 3000 probe sets (representing about 2500 genes) that passed a variance filter were used in this analysis. We applied the algorithm on each of the ALL subtypes separately, in order to avoid 'inter-subtype' noise. ALL subtypes with large numbers of samples were the first to be analyzed. We used both unsupervised and supervised approaches. The search for the characteristic gene set was performed in a totally "blind" way, without posing any hypothesis regarding either the separating gene set or the resulting partition of the samples.

Surprisingly, we observed two unanticipated partitions of the ALL patients to novel subgroups, induced by two different clusters of genes (referred to as *cluster no.1* and *cluster no. 2*). In both partitions clear and sharp separation was demonstrated, but while for *cluster no.2* the biological meaning of this separation is still vague, we suggest an interesting interpretation to the partition induced by the expression levels of the genes of cluster no. 1.

2.1 Class discovery within the TEL-AML1 subtype

Cluster no. 1 was obtained while starting the class discovery procedure with the TEL-AML1 subtype.

2.1.1 Discovery of novel subgroup – unsupervised step

When we applied the CTWC algorithm on the TEL-AML1 samples, we noticed an interesting group of 16 probe-sets (Table 3) that separated the TEL-AML1 subtype very clearly into two sub-groups (Figure 8): in 8 TEL-AML1 samples

these probe-sets had high expression levels whereas in the remaining 71 samples their expression levels were relatively low (see Appendix, Table 1). The distinct group of 8 samples shared no clinical label (such as same protocol of treatment or same prognosis). Many of the differentiating genes belong to the JAK/STAT1 pathway induced by interferon.

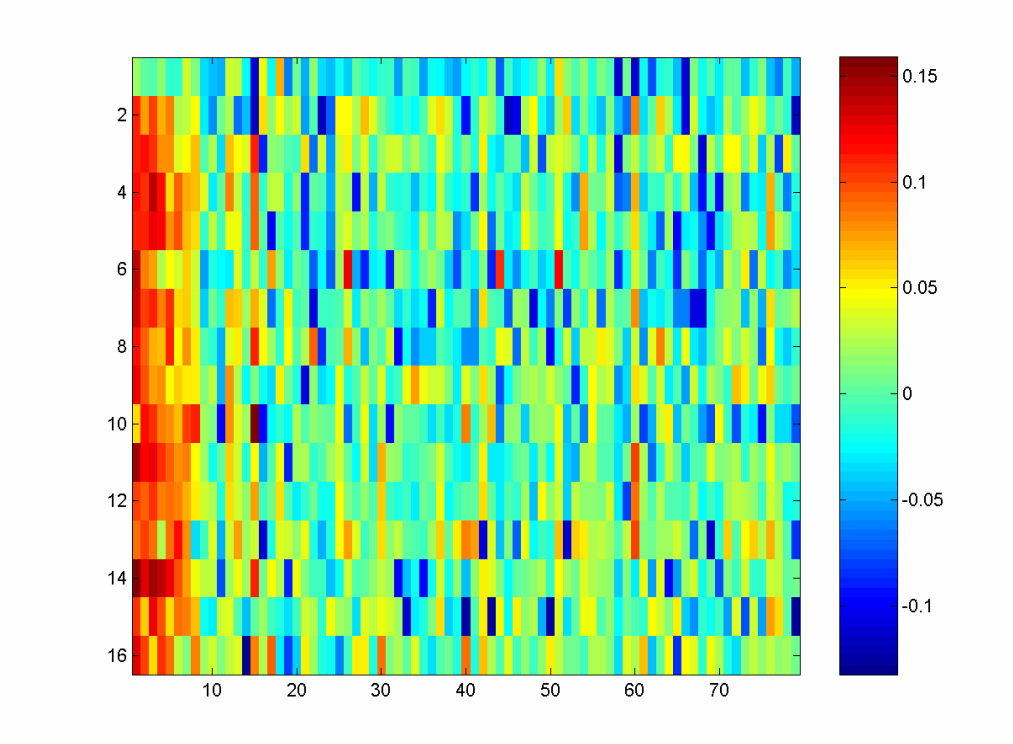


Figure 8 | Expression values of the cluster of 16 genes, found by CTWC, in 79 *TEL-AML1* samples. The values are centered (mean of each gene = 0) and normalized (STD = 1). These genes are overexpressed in a group of 8 (out of 79) TEL-AML1 samples.

2.1.2 Refine the list of genes – supervised step

Next, we refined the list of these genes, using supervised analysis (see methods section). First, we took the separation into the two groups of 8 versus 71 samples as "ground truth" and searched for genes that differentiate between these two groups. This search was performed on an extended set of 6500 genes. 184 probe-sets passed the TNoM as differentiating, with p-values below 0.05. To overcome the problem of multiple comparisons we applied the FDR method; 23 were identified as separating at an FDR level of 5% (Table 3).

The practical meaning of this statement is that out of these 23 probe-sets we expect about one to be a false positive, present due to random fluctuations.

2.1.3 Expanding the novel subgroup – unsupervised step

The next step of our iterative refinement process was unsupervised; we used the expression levels of the 23 probe-sets found above to characterize all samples, and clustered them using SPC. This way we identified a group of 50 samples, selected from all the ALL subtypes that have high expression levels of the 23 probe-sets (Figure 9 and Figure 10). The average fold change of the absolute expression values between the two sub-groups is 2.63 (Table 2). This group of samples consists mainly of *hyperdiploid>50* and *TEL-AML1* subtypes, but contains almost all other subtypes as well (Table 4). The *hyperdiploid>50* subtype was significantly over-represented among the 50 samples with high expression; no other clinical label, specific to these samples, was found.

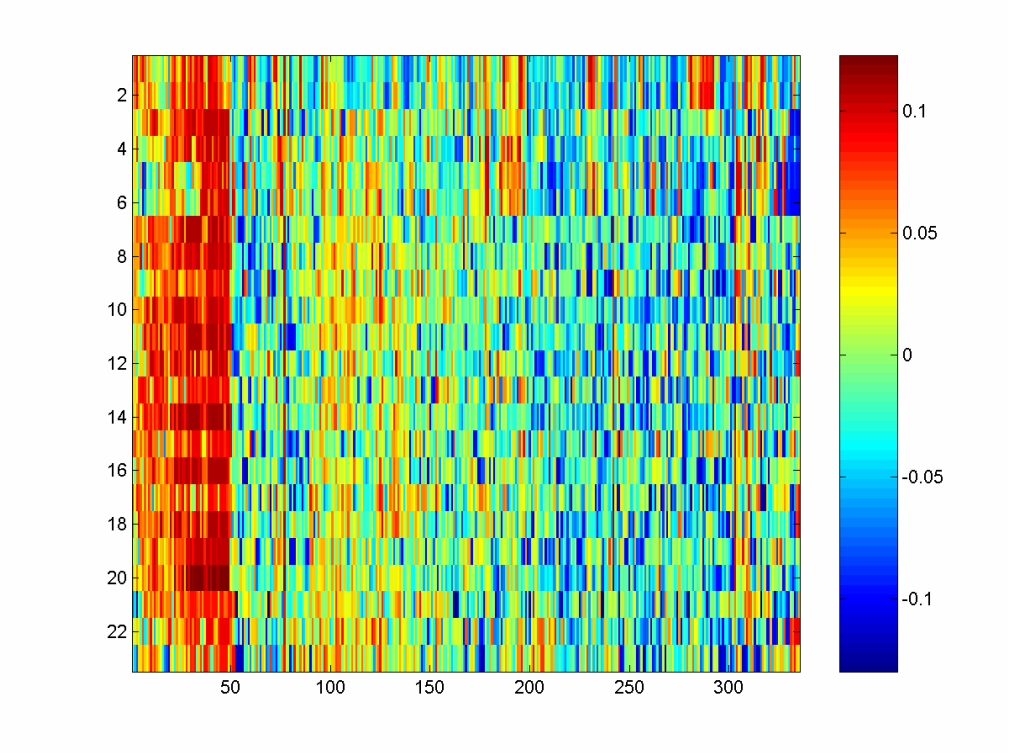


Figure 9 | Expression values of the 23 genes in 335 ALL samples. The values are centered (mean of each gene = 0) and normalized (STD = 1). These genes are correlated and overexpressed in a group of 50 ALL samples, that consist mainly of *hyperdiploid>50* and *TEL-AML1* subtypes.

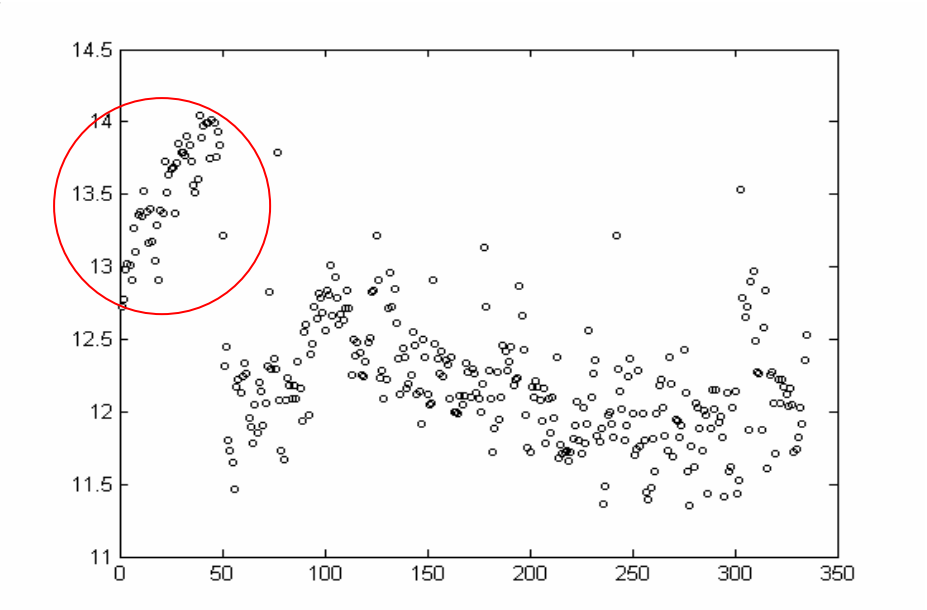


Figure 10 | The mean absolute expression values of all 23 probe-sets for each of the 335 samples. The 50 samples of cluster no. 1 are circled in red. The values are log scaled.

Table 2| The novel sub-groups' means of the absolute expression values of each probe-set. The last column represents the fold change of each probe-set between the sub-groups. The average fold change is 2.63.

Probe-set ID*	Samples 1-50	Samples 51-335	Fold change
38014_at	24100	16884	1.4273
1358_s_at	4870	2215	2.1964
1107_s_at	21260	7626	2.7875
36412_s_at	13650	4030	3.3875
38432_at	11600	4196	2.7648
37641_at	8680	2305	3.7646
38662_at	6640	3309	2.0061
41745_at	77170	42048	1.8352
37014_at	43140	8707	4.9547
37353_g_at	4910	3427	1.432
37360_at	13950	7366	1.8934
1184_at	22220	18058	1.2306
915_at	6880	1349	5.0982
38584_at	7470	2171	3.441
39263_at	13220	6328	2.0893
39264_at	3420	1576	2.1708
36927_at	17160	2122	8.0905
40505_at	16490	10999	1.4991
41171_at	30870	25571	1.2073
37352_at	2340	1646	1.4217
38389_at	2920	1071	2.7268
676_g_at	146260	89327	1.6374
464_s_at	7800	4953	1.5746

* See Table 3 for additional gene annotation information

2.1.4 Refine the final group of genes – supervised step

To complete the refinement process, supervised analysis was performed again, using TNoM on 6500 probe-sets, revealing 28 probe-sets that most significantly separate the new sub-group of 50 samples from the remaining 285 samples.

Table 3 | Genes that separate the ALL samples into two subgroups

Gene Probe ID	Title	Gene Symbol	TEL-AML1 CTWC	TNoM 1 (P val)	TNoM 2 (P val)
36927_at	chromosome 1 open reading frame 29	C1orf29	+	0.000115	6.69E-35
925_at	**interferon, gamma-inducible protein 30**	IFI30	+		2.51E-32
915_at (32814_at)	**interferon-induced protein with tetratricopeptide repeats 1**	IFIT1	+	6.06E-06	2.51E-32
37641_at	**interferon-induced protein 44**	IFI44	+	6.06E-06	4.44E-31
37014_at	**myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)**	MX1	+	6.06E-06	7.40E-30
38584_at	**interferon-induced protein with tetratricopeptide repeats 4**	IFIT4	+	0.000115	1.17E-28
1107_s_at (38432_at)	**interferon, alpha-inducible protein (clone IFI-15K)**	G1P2	+	6.06E-09	3.41E-25
38389_at	**2',5'-oligoadenylate synthetase 1, 40/46kDa**	OAS1	+	0.000115	4.43E-24
39263_at (39264_at)	**2'-5'-oligoadenylate synthetase 2, 69/71kDa**	OAS2	+	0.000115	5.51E-23
38014_at	**adenosine deaminase, RNA-specific**	ADAR		6.06E-09	6.57E-22
38517_at	**interferon-stimulated transcription factor 3, gamma 48kDa**	ISGF3G			8.76E-19
1358_s_at			+	6.06E-09	8.35E-16
38662_at	Homo sapiens, clone IMAGE:4074138, mRNA sequence			6.06E-06	7.67E-15
37360_at	**lymphocyte antigen 6 complex, locus E**	LY6E		6.06E-06	3.94E-11
35718_at	**SP110 nuclear body protein**	SP110			2.35E-09
33339_g_at (32860_g_at)	**signal transducer and activator of transcription 1, 91kDa**	STAT1	+		1.74E-08
464_s_at	**interferon-induced protein 35**	IFI35		0.000115	8.77E-07
914_g_at (36383_at)	v-ets erythroblastosis virus E26 oncogene like (avian)	ERG			5.98E-06
40054_at	KIAA0082 protein	KIAA0082			5.98E-06
37352_at (3753_g_at)	**nuclear antigen Sp100**	SP100	+	6.06E-06	5.98E-06
34947_at (41472_at)	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G	APOBEC3G			3.97E-05
36845_at	nuclear matrix protein NXP-2	NXP-2			3.97E-05
40505_at	ubiquitin-conjugating enzyme E2L 6	UBE2L6		0.000115	3.97E-05
41841_at	Homo sapiens clone 23718 mRNA sequence				0.000257
39061_at	bone marrow stromal cell antigen 2	BST2			0.000257
32800_at	retinoid X receptor, alpha	RXRA			0.000257
38805_at	TGFB-induced factor (TALE family homeobox)	TGIF			0.000257

890_at	ubiquitin-conjugating enzyme E2A (RAD6 homolog)	UBE2A			0.000257
40852_at	tudor repeat associator with PCTAIRE 2	PCTAIRE2BP	+		
32775_r_at	phospholipid scramblase 1	PLSCR1	+		
36412_s_at	**interferon regulatory factor 7**	IRF7	+	2.36E-07	
41745_at	**interferon induced transmembrane protein 3 (1-8U)**	IFITM3		6.06E-06	
676_g_at	6-pyruvoyltetrahydropterin synthase	PTS		0.000115	
1184_at (41171_at)	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)	PSME2		6.06E-06	

We list in the 'TEL-AML CTWC' column the genes that were obtained by the initial CTWC analysis; two out of the 16 probe-sets corresponded to the same gene and hence only 15 genes are marked. The 'TNoM1' column gives P-value of the genes that separate 8 versus 71 *TEL-AML1* samples according to the TNoM test, at FDR=0.05. Only 19 genes are indicated (out of 23 probe-sets), again because of multiple representation.. The 'TNoM2' column indicates 28 genes that separate all the ALLs into two subgroups of 50 versus 285 samples (see text). The genes that are known to be part of the interferon-JAK/STAT pathway are in **bold face**. *In cases when two probe sets represent the same gene symbol, the lower p-value was taken.*

Table 4 | The number of samples from each subtype in the Yeoh et al.[7] dataset.

Subtype name	**Number of samples**	**Number in subgroup**
Hyperdip>50	65	24
TEL-AML1	79	10
Pseudodip	29	4
Normal	19	4
Hyperdip 47-50	23	3
T-ALL	45	2
BCR-ABL	16	2
MLL	21	1
Hypodip	11	0
E2A-PBX1	27	0
Total:	**335**	**50**
Novel subtype (as found by Yeoh et al.)	14	6

2.1.5 Analyzing other datasets[34-36,38-45]

We now turned to search for other types of cancer in which a similar finding may hold; we tested whether we can find a sub-division of samples in other datasets on the basis of the expression levels of genes from the same pathway.

However, in each of the following datasets we had to use a different subset of the separating genes: In some cases some of the relevant genes did not appear in the dataset (different versions of DNA chips were used) and in other cases some genes had too many missing values. We ran the SPC algorithm for each of the datasets, using for each the appropriate subset of our gene list. Our aim was to find a distinct group of samples, in which these genes were overexpressed. In addition, we checked the samples' labels, in order to find common clinical indicators, shared by the members of the selected subgroup.

Firstly, we analyzed the leukemia data of Golub et al.[34], Armstrong et al.[35] and Rozovskia et al.[36]. In each of these datasets we also found clear subgroups (containing about 10% of the samples), with overexpressed levels of these genes. Again, no common label was shared by the subgroup's members.

Next, we ran CTWC with datasets of other types of cancer : lymphoma[40], prostate[41,44] mix of cancers[39,42], ovary[43], lung[38] and breast[45]. In the lymphoma and prostate data sets we found very small or negligible sub-groups of samples that co-expressed our sub-group of genes. Analysis of the data published by Ramaswamy et al.[39], which contains samples from various types of cancer, revealed a small sub-group, 7 out of 280, that also contains samples from other types of cancer, but mainly from leukemia, lymphoma and even from normal peripheral blood samples. In the lung cancer data set of Bhattacharjee et al.[38], a subset of 1.5% of the samples showed high expression levels. Another 20% of the samples exhibited intermediate expression levels of these genes.

In two datasets we did find significant overexpression: ovary and breast cancers. In the ovary cancer data set of Welsh et al.[43] the interferon-related genes were overexpressed in about 20% of the samples and in the breast cancer data of van 't Veer et al.[45]overexpression in about 40% of the samples was found (Figure 11) .

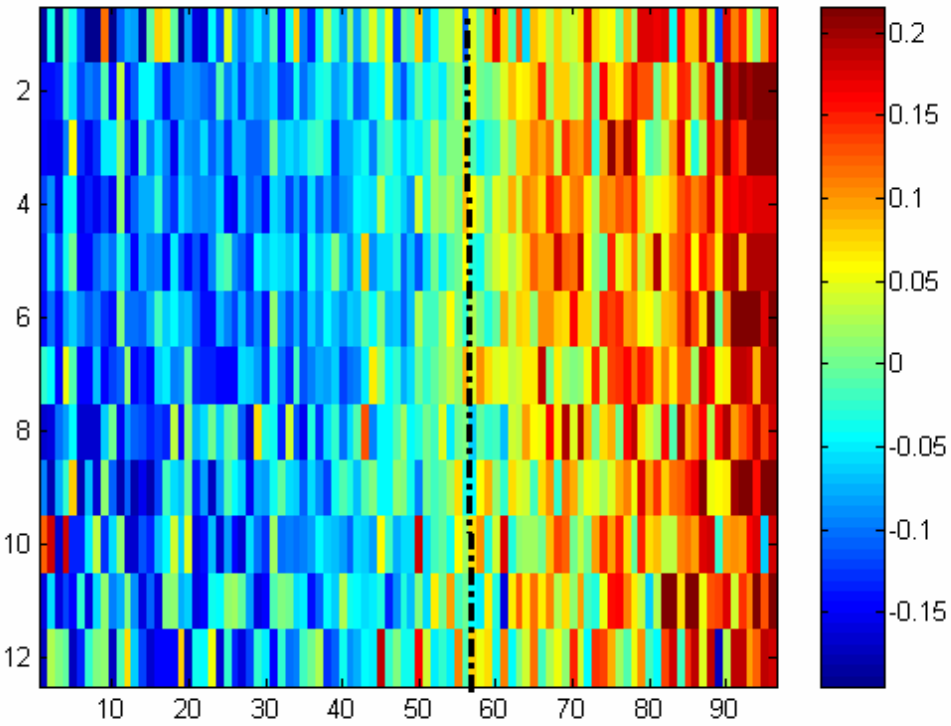


Figure 11| Expression values of the 12 probe sets in 96 breast cancer samples[45]. The values are centered (mean of each gene = 0) and normalized (STD = 1). Approximately 40% of the samples (to the right of the black dotted line) overexpress these genes. The 12 probe-sets correspond to 11 differentiating genes: ifi35, stat1, ly6e, oas2, oas1, ifit1, ube2l6, ifit4, plscr1, irf7, mx1

2.1.6 Cluster No. 1 - Discussion

We observed cluster no. 1 while performing a *class discovery* procedure with published gene expression data of childhood ALL's of the TEL-AML1 subtype. Surprisingly, a special set of genes was found to be highly expressed in a small minority (0-14%) of samples of the various leukemia subtypes, in a relatively high fraction (37%) of the hyperdiploid (>50) ALL subgroup and in a clear subgroup of the TEL-AML1 ALL

2.1.6.1 Differentiating genes - biological meaning

A set of 34 genes was obtained during all the supervised and the unsupervised steps. These genes separate the TEL-AML1 samples or the whole group of the ALL samples into two subgroups, on the basis of their expression levels. It

turns out that 29 of the 34 genes that appear in the gene cluster are known; 17 of these are directly related to the **interferon-JAK/STAT1** pathway (Table 3). These include signal transducer and activator of transcription 1 (STAT1) and interferon regulatory factor 7 (IRF7), both involved in signal transduction downstream to interferon receptors, as well as many interferon alpha induced proteins such as interferon induced protein 44, interferon induced transmembrane protein 3, interferon induced protein 35, 2'5'-oligoadenylate synthetase 1 and 2, myxovirus resistance 1 interferon-inducible protein 78 and adenosine deaminase RNA specific. Interferon gamma-induced proteins such as protein 30 and interferon gamma-induced transcription factor 3 were also found in the special gene cluster. Interestingly, several ubiquitin-conjugating enzymes such as E2L6 and E2A and proteasome system components such as activator subunit 2 (PA28 beta), some of them known to be induced by interferon, were also expressed in the cluster. Such proteins are involved in the generation of antigenic peptides that are presented to CD8 T cells by MHC class I molecules. Taken together, many genes relevant to the immune response, and in particular immune response to viruses, were found to be expressed in the special cluster of genes that are highly expressed in the hyperdiploid leukemia and TEL-AML1 variants. Of great interest is the finding of the expression of apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G (APOBEC3G) in the gene cluster. This enzyme was shown lately[47,48] to confer antiretroviral defense against HIV and other retroviruses through lethal editing of nascent reverse transcripts. Hypermutation by editing mediated by this enzyme was shown to be an innate defense mechanism against retroviruses. One may speculate that the expression of this gene is an indication for retrovirus involvement in childhood leukemogenesis.

2.1.6.2 Subtypes of ALL within cluster no.1

The JAK/STAT genes were overexpressed in samples of various leukemia subtypes, but particularly in a relatively high fraction (37%) of the hyperdiploid (>50) ALL subgroup and in a clear representation of the TEL-AML1 ALL subgroup (Table 4); it turns out that these two subtypes have a common feature: both constitute a large part of the cases in the early childhood peak of leukemia (See introduction section, Figure 2).

2.1.6.3 Inferring from other datasets' analyses

Search for the interferon gene cluster in other datasets of several malignant diseases indicates that in other datasets of leukemia about 10% of the samples expressed the special gene set. In lymphomas, prostate cancer and datasets of a mixture of tumors none or very low percentage of the samples expressed the set. In the datasets of 49 ovarian cancer samples and 203 lung cancer samples, subsets of ~20% expressed the gene set. In the dataset of breast cancer[45], a subset of ~40% of the samples overexpressed the genes. It is of great interest that some epidemiologic studies suggested a role for infection in the pathogenesis of cancers other than leukemia; among them ovarian[49], lung[50], and breast[51] tumors. The finding that about 40% of breast cancer samples expressed the 'infection associated' gene signature is of special significance. Retrovirus-like particles were demonstrated in a breast cancer cell line[52] and Mouse mammary tumor virus (MMTV)-like gene sequences were detected by PCR in about 40% of breast cancer samples in several studies[51,53]. Interestingly, the percentage of cases where retroviral gene sequences were identified is very similar to the percentage of cases where the interferon gene signature was identified (~ 40%). The experiment to be done is to look at the same tumor samples for both interferon-associated gene expression and for the MMTV-like gene sequences.

The prominent appearance of hyperdiploid leukemic samples (that occur at early childhood, where viral infection is most natural to occur) and the overexpression of these genes in cancers that are suspected to have a role for infection in the their pathogenesis, strengthens the hypothesis of Greaves and Kinlen[17,21].

The samples with high expression of the special genes constitute a small minority of the leukemic samples. Even in the hyperdiploid subgroup only one third of the samples were positive. Several explanations can be suggested for this finding. First, infection can represent only one type of "second hit" in childhood leukemia, and other mechanisms may operate in the rest of the cases. Second, the role of infection may be indirect, via the dysregulated

immune response, as suggested by the Greaves hypothesis. Under such a scenario the infectious agent can contribute to leukemogenesis in a transient "hit and run" fashion and its fingerprints may not be found at the time of leukemia diagnosis.

Our finding, of a highly expressed "interferon cluster", combined with the epidemiological evidence, implies that a viral infection induced an immune response, which resulted in interferon secretion, activation of interferon receptors and JAK/STAT signaling, resulting in the activation of many interferon regulated genes. Nevertheless, a less likely possibility, that the pathway was activated by a mutation, independent of viral infection, cannot be completely ruled out. The role of aberrant STAT signaling and constitutive STAT activation in leukemia is the subject of recent intensive research[54]. Activation of STATs and specifically STAT1 has been demonstrated in leukemic cell lines and blasts from 22-100% of patients with AML by several groups[54]. There are several reports of constitutive STAT1 activity in some of the ALL samples, similar to our findings. It is unclear whether constitutive STAT activation itself is the cause or the result of a transforming process.

Our findings identified a set of interferon regulated genes suggestive of immune response to viral infection mainly in the hyperdiploid leukemia subtype. Taken together with the epidemiology-based hypotheses of Greaves and Kinlen, suggesting abnormal activation of immune response as contributing factor in leukemogenesis, the possible role of infectious agents, and in particular viruses, is strengthened. We also showed overexpression level of the anti-viral genes in ovary and breast cancers samples; other epidemiologic studies suggested a role for infection in the pathogenesis these cancers. These findings might strengthen the above hypotheses.
A search of infectious agents, mainly in those leukemias where high expression of the "interferon cluster" is identified, may lead to the identification of the putative viral pathogen. The implications of the finding of such a causative agent on diagnosis, therapy and prevention of childhood leukemia are clear.

2.2 Class discovery within the hyperdiploid>50 subtype

Cluster no. 2 was obtained while starting the class discovery procedure with the hyperdiploid>50 subtype.

2.2.1 Discovery of another subgroup – unsupervised step

Applying the CTWC algorithm on the 65 samples of *hyperdiploid>50* subtype revealed an interesting group of 14 probe-sets (Table 5) that separated the *hyperdiploid* subtype sharply into two sub-groups (Figure 12): in 15 hyperdiploid>50 these probe-sets had relatively low expression levels whereas in the remaining 50 samples their expression levels were relatively high (see Appendix, Table 2). As before, the distinct group of 15 samples shared no known clinical label.

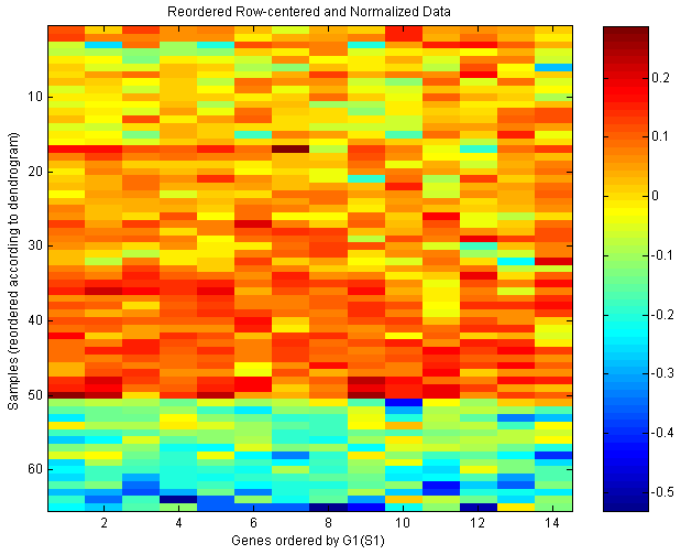


Figure 12 | Expression values of the cluster of 14 genes, found by CTWC, in 65 *hyperdiploid>50* samples. The values are centered (mean of each genes = 0) and normalized (STD = 1). These genes are underexpressed in a cluster of 15 (out of 65) *hyperdiploid>50* samples. Columns represent the genes and rows represent the samples.

2.2.2 Refine the list of genes – supervised step

Next, we refined the list of these genes, using the same approach that we used while analyzing *cluster no 1*. We searched for genes that differentiate between these two groups, of 15 versus 50 samples, using TNoM. This search was

performed on an extended set of 6500 genes. 38 probe-sets were identified as separating at an FDR level of 1% (Figure 13 and Table 5).

2.2.3 Expanding the novel subgroup – unsupervised step

Subsequently, we used the expression levels of the 38 probe-sets found above to characterize all 335 ALL samples, and clustered them using SPC. This way we identified a group of 37 samples, selected from all the ALL subtypes that have high expression levels of the 38 probe-sets (Figure 14). This group of samples contains almost all subtypes; No common translocation or other clinical label, specific to these samples, were found (see Appendix, Table 3).

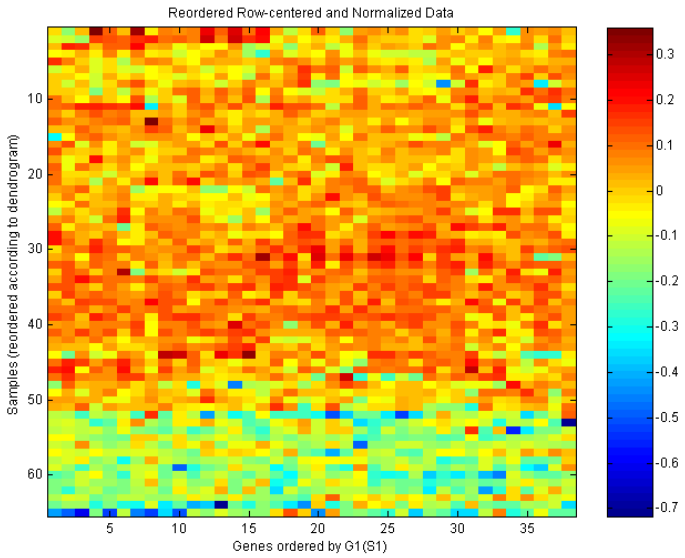


Figure 13| Expression values of the expanded group of 38 genes, found by TNoM, in 65 *hyperdiploid>50* samples. The samples are sorted according to the CTWC results of the previous step. The 38 genes are underexpressed in a cluster of 15 (out of 65) *hyperdiploid>50* samples. Columns represent the genes and rows represent the samples.

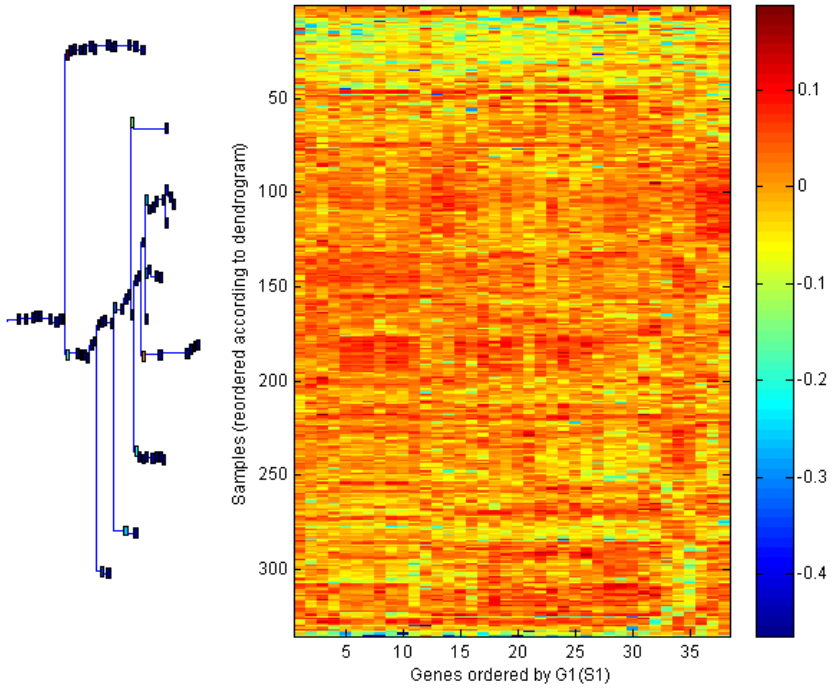


Figure 14| Expression values of the expanded group of 38 genes. The 38 genes are underexpressed in a cluster of 37 (out of 335) samples. Columns represent the genes and rows represent the samples.

Table 5 | Genes that separate the hyperdiploid > 50 subtype into two novel subgroups

Probe Set	Title	Gene symbol ↓	Hyperdiploid CTWC
1474_s_at			+
1839_at			
1903_at			
1928_s_at			
31608_g_at			
41842_at	hypothetical gene supported by AK026880		
262_at 263_g_at 36684_at	adenosylmethionine decarboxylase 1	AMD1	
1984_s_at	Rho GDP dissociation inhibitor (GDI) beta	ARHGDIB	+
40096_at	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha	ATP5A1	
38085_at	chromobox homolog 3 (HP1 gamma homolog, Drosophila)	CBX3	
38791_at	dolichyl-diphosphooligosaccharide-protein	DDOST	

	glycosyltransferase		
826_at	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	DDX3X	
33647_s_at	GM2 ganglioside activator protein	GM2A	
1139_at 33635_at	guanine nucleotide binding protein (G protein), alpha 13	GNA13	+
466_at	general transcription factor II, i	GTF2I	
36783_f_at	Krueppel-related zinc finger protein	H-plk	
41300_s_at	integral membrane protein 2B	ITM2B	
1457_at	Janus kinase 1 (a protein tyrosine kinase)	JAK1	+
39471_at	membrane component, chromosome 11, surface marker 1	M11S1	
32571_at	methionine adenosyltransferase II, alpha	MAT2A	
37711_at	MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)	MEF2C	
1472_g_at 1473_s_at 1475_s_at 1476_s_at	v-myb myeloblastosis viral oncogene homolog (avian)	MYB	+ + + +
38614_s_at	O-linked N-acetylglucosamine (GlcNAc) transferase	OGT	
40440_at	PAI-1 mRNA-binding protein	PAI-RBP1	
1318_at	retinoblastoma binding protein 4	RBBP4	+
865_at	ribosomal protein S6 kinase, 90kDa, polypeptide 3	RPS6KA3	
351_f_at	splicing factor, arginine/serine-rich 3	SFRS3	+
34709_r_at	stromal antigen 2	STAG2	
32548_at	unactive progesterone receptor, 23 kD	TEBP	
1581_s_at	topoisomerase (DNA) II beta 180kDa	TOP2B	+
504_at	ubiquitin-conjugating enzyme E2D 3 (UBC4/5 homolog, yeast)	UBE2D3	+
34642_at	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide	YWHAZ	+
826_at	*Homo sapiens helicase like protein 2*	*DDX14*	+

We list in the 'hyperdiploid CTWC' column the genes that were obtained by the initial CTWC analysis; DDX14 (*in italic face*) was obtained by the first unsupervised analysis but didn't separate significantly the 335 samples into two sub-groups.

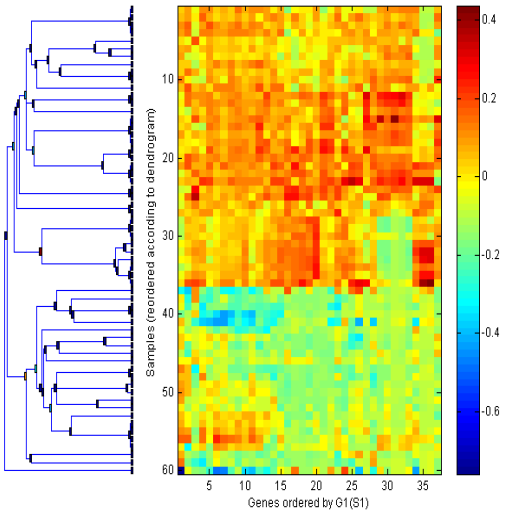
2.2.4 Analyzing other datasets[34-36,39-41]

Since we can not point to a common pathway for the differentiating genes of *cluster no. 2*, it is vital to test whether we can find a similar sub-division in other leukemia datasets and in other types of cancer. For this end, we used the 38 separating genes from the refined list which was obtained by supervised analysis of the hyperdiploid>50 subgroup. In each of the following datasets we had to use a different subset of the 38 separating genes, since some genes simply did not appear in these datasets and others had too many missing values. We ran the SPC algorithm for each of the data, using the appropriate sub-set of our gene list. In the cases in which we indeed found a similar sub-division, we checked the labels shared by the member of the selected subgroup. A common label for the similar sub-division in other datasets might reveal the meaning of the novel sub-group of ALL samples, and shed a light on the function of the differentiating genes.

2.2.4.1 Leukemia datasets

The analyses of the leukemia datasets, published by Golub et al[34] and Armstrong et al[35]., revealed small sub-groups that exhibited relatively low levels of mRNA expression. We find among these samples different subtypes of leukemia. No other common clinical label was shared.

Clear and interesting results were obtained while analyzing the expression data of Rozovskaia et al[36] (Figure 15). We observed very sharp separation; in 40% of the samples the genes were underexpressed, while in the remaining 60% they were overexpressed. The distribution of the leukemia's subtypes between these two subgroups is not random at all (see Appendix, Table 4): The first subgroup (overexpression) consists of MLL t(4:11), CD10- ,AML(*:11) and the normal samples. The second subgroup consists of ALL controls, AML control, AML-dup and 3 MLL t(4:11) samples. These 3 MLL samples were found to have lower expression levels of "differentiation sensitive genes"[36] than the other MLLs. The normal samples overexpressed all but one of the 37 genes; the C-MYB gene (which is represented by a few probe-sets).



MLLt(4:11), CD10-,AML(*:11)
Normal
ALL controls, AML control, AML-dup

Figure 15 | Clustering the 60 samples of Rozovskaia et al. based on 37 genes from *cluster no. 2*. Sharp and clear separation was detected. The samples' labels are not randomly distributed. Columns represent the genes and rows represent the samples.

2.2.4.2 Analyzing other datasets[39-41]

In the dataset of Ramaswamy et al., the samples were collected from different types of cancer. Analysis of this data, using genes from *cluster no. 2*, showed that almost all of the 60 blood-related samples had relatively high expression levels. The remaining 220 samples, which were produced from non-blood tissues, exhibited relatively low expression levels. In the prostate data of Singh et al. 102 samples were collected from normal and tumor tissues. We observed two distinct pattern of expression in two different sub-groups of samples. Each sub-group consists of both normal and tumor prostate tissues. In the lymphoma data of Shipp et al. we observed about 15 samples (out of 77) that underexpressed these genes; the genes from cluster no. 2 were co-expressed in all of these datasets.

2.2.5 Search for common transcription factors (TF)

A further analysis was needed, since the biological meaning of cluster no. 2 remained vague at this point. This analysis should involve information from additional sources, beyond expression data. We decided to examine the promoter regions of the genes in this cluster, in order to identify the main transcriptional regulators and to reveal the underlying regulatory network. To

this aim we used recently developed program called POC. This program is aimed at finding significant TFs' binding-sites in a given cluster's promoters (for more details see [55]).

We used the promoters of 20 out of the 38 probe-sets, due to multiple representation of some genes and the lack of known promoters of others. The promoter region was defined as the 1000 upstream base-pairs, from the start codon, Using POC, we have screened the promoters of the genes in this cluster for all known human regulatory motifs in the TRANSFAC database (http://www.gene-regulation.com/). The POC program identified four transcription factors, Hoxa5 Hoxa9, Hox-7 and COMP1, which were overrepresented in the cluster's promoters. The first three belong to the HOX family which contains genes that encode a set of master transcription factors which act during development to control pattern formation, differentiation and proliferation.

2.2.6 Cluster No. 2 – Discussion

Cluster no. 2 was identified while performing a *class discovery* procedure on the *hyperdiploid>50* subtype samples of the data of Yeoh et al[7]. Contrary to findings in class no 1, the co-expressed genes here share no common known pathway. These genes demonstrated interesting expression patterns in several distinct datasets. In addition to the data of Yeoh et al., three other leukemia datasets were examined. In each one of them, a clear separation of the samples was demonstrated, using the genes of *cluster no 2*. The genes were again co-expressed in the lymphoma and prostate, while separating the samples of these data-sets into novel sub-groups. The distribution of the different subtypes of leukemia within the new cluster provided no new insights in all but one dataset; a non-random distribution of the samples was obtained within the two clusters formed in the leukemia data of Rozovskaia et al[36]**.** One cluster consists mainly of samples that are thought to be immature and the other consists of samples that are supposed to be late differentiated. The composition of the obtained clusters could imply that these are differentiation sensitive genes, but the knowledge we currently hold for these genes does not support this

hypothesis. Moreover, comparison of our list of genes with previously published differentiation sensitive genes[36] yields no significant overlap.

The extremely preliminary and basic recent analysis of the promoter regions yielded interesting findings. This analysis identified four transcription factors, Hoxa5 Hoxa9, Hox-7 and COMP1, which were overrepresented in the cluster's promoters. Three of these are part of the HOX family and are known to be related to leukemia. Gene expression profiling revealed that expression of the fusion protein MLL-AF9 led to over-expression of Hoxa5, Hoxa6, Hoxa7, Hoxa9 and Hoxa10[56] In an experimental setting, a high proportion of mice transplanted with bone marrow cells overexpressing Hoxa9 developed AML[57]. In humans, the MLL/ALL-1 gene, which normally functions to maintain Hox gene expression, is a frequent target of chromosomal rearrangements. Joh *et al*[58] demonstrated that a chimeric MLL-LTG9 protein led to the inhibition of Hoxa7, Hoxb7 and Hoxc9 expression in mouse myeloid cells. Rozovskaia *et al*[59] found that the t(4;11) translocation was associated with increased expression of HOXA9. Importantly, HOXA9 has been identified in a gene expression array-based screen as the single gene whose expression most correlated with treatment failure in AML[34].

The strong association of the HOX family and leukemia allow us to suggest several different hypotheses, in order to elucidate the correlation between the above Hox TFs, the genes of cluster no. 2 and the novel subgroup of leukemia patients. These hypotheses should be examined deeply and carefully, but unfortunately such an examination is not in the scope of this thesis.

Nevertheless, we do present these findings, due to the unusually clear separation we obtained within one leukemia subtype. It is a reasonable assumption that these genes are co-regulated and share a certain pathway with specific function. The distinct two sub-groups may be an evidence for some important biological phenomenon. At the moment we lack essential information, such as additional gene annotations or supplementary clinical data for the samples to reveal the biological process. Analysis of more advanced chips can be useful too, due to the constant improvement of the DNA chips

technology. More thorough examination of the promoters regions and careful exploration of the HOX TFs are likely to elucidate the hidden connections between the genes. Analysis of other published datasets, containing more relevant clinical labels for the samples, can illuminate the biology behind this obscure pathway.

We believe that in the near future, with the help of new data, the yet unrecognized pathway, whose existence we demonstrated, will be elucidated and completed and its role in the pathogenesis of leukemia will become clear.

3 Materials and Methods

3.1 Patients

There are several publicly available gene expression datasets on leukaemia[7,34-36] We decided to start our analysis with the data of Yeoh et al.[7] since it had the largest number of patients. The experiments were done using Affymetrix U95A chips containing 12,533 probe sets. Expression levels were measured for 335 samples, of bone marrow and peripheral blood, representing several different ALL subtypes (*T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL, hyperdiploid >50* chromosomes*, hyperdiploid 47-50* and *hypodiploid*). We also used additional details, such as protocol of treatment and prognosis that were supplied by the authors.

To substantiate our findings that were derived from the data of Yeoh et al.[7], we analyzed other publicly available datasets of leukemia and of other cancers[34-36,38-45].

3.2 Preprocessing and filtering

We started out with an expression matrix organized in 335 columns (samples) and 12,533 rows (genes). Each value in the matrix is the gene expression level of a certain gene for one patient. First, rows (genes) in which more than 20% of the values were lower than some threshold (we chose T=10) were removed. After this filtering we were left with 6,653 genes. In these rows the values that were lower than T were replaced by estimates based on the values of the 13 nearest neighbors' genes[60]. Next, logarithm (base 2) of each entry was taken, and the genes were filtered on the basis of their variation across the samples. Two sets, of 3000 and of 6500 genes, were chosen for the CTWC step and the TNoM test respectively, based on their standard deviation.

3.3 Unsupervised Analysis

3.3.1 Clustering Analysis

In order to separate the ALL samples into **unanticipated sub-groups**, we search for a cluster (e.g. correlated set) of genes with a distinct expression profile in one part of the samples, and another profile in the other part. This task can be done using an unsupervised clustering algorithm. Standard statistical methods are useful for *hypothesis testing:* in the present context, if we know that there are two subclasses of the disease, A and B, we can use standard methods in order to find the genes whose expression levels differentiate these two classes. Hypothesis testing will not reveal unexpected partitions; unsupervised techniques, such as clustering, are more suited for such a task. Once a class is proposed, a hypothesis can be formulated and the powerful standard methods of statistical analysis can be used to substantiate and validate the finding. We used here such a hybrid approach, combining unsupervised cluster analysis with standard supervised statistical tests.

3.3.2 Super Paramagnetic Clustering (SPC)

We use SPC[61] as our clustering algorithm. SPC is based on the physical properties of an inhomogeneous ferromagnet[61,62] ; it has been tested on data from a large number of problem areas including image analysis and speech recognition, computer vision [CCP] and gene expression [SPC on gene exp]. SPC belongs to the family of hierarchical clustering algorithms. Its main advantage is that it provides for each cluster of the hierarchy a stability *index,* whose value indicates the statistical significance or reliability of the cluster. This allows us to recognize stable clusters, an essential ingredient for using CTWC. Other attractive features of SPC are stability against noise in the data and that one does not need to specify in advance the number of clusters. For more details see [61,62].

3.3.3 Coupled Two-Way Clustering (CTWC)

Coupled Two-Way Clustering[63] is a method for reducing noise by focusing on small subsets of genes and samples. This is achieved by using only genes (and samples) that were identified previously as a stable cluster for the clustering

process. The procedure is iterative; each stable cluster of genes that was found is used, in the next iteration, for the clustering of each of the stable clusters of samples that were previously found and vice versa. This process is repeated until no more clusters answering our criteria are discovered.

This method is especially appropriate for analyzing DNA microarray data, where different biological processes are simultaneously occurring, influencing different genes and samples. As a result, a great deal of noise is present, masking the effects individual processes. By focusing on correlated groups of genes and samples, CTWC is able to minimize the noise generated by the majority of genes and identify specific biological processes involving specific genes or samples. Moreover, by using a group of correlated genes, noise of the individual measurements averages out and is reduced.

CTWC can be used with a variety of clustering algorithms. We use CTWC with SPC because of its robustness against noise and because it is one of the few algorithms that provides a reliable stability index to each cluster.

3.4 Supervised Analysis

We used supervised methods in order to expand and refine the lists of genes that were obtained in the unsupervised CTWC analysis. Using hypothesis testing (TNoM[64]), we identified genes, one at a time, whose expression differed between the two groups of samples that were identified by CTWC: the group in which the genes (initially found by clustering) were overexpressed, and all the remaining samples. The process was then repeated: an extended group of separating genes was then used to identify, in a supervised manner, samples that belong to groups of relatively high expression. This procedure is reminiscent in spirit of the signature method[46].

3.4.1 Parametric vs. nonparametric methods

Parametric approaches model expression profiles within a parametric representation, and ask how different the parameters of groups *A* and *B* are[65].

The *t test*, for instance, compares the mean of group *A* and the mean of group *B*. The t test assumes approximately normal distribution and roughly similar variances.

On the opposite side, nonparametric methods (such as TNoM) use no *a priori* assumptions, meaning no assumptions are made about the distribution of expression profiles in the data. Instead, we attempt to directly examine the degree to which the two groups, A and B, are distinguished. The main weakness of parametric approaches is the assumptions that they make about the data. For example, outliers can significantly bias the *t test* score by changing the variance estimated in the samples. In a similar manner, working in logarithmic scale can have drastic impact on the scores. Nonparametric approaches are more robust against these types of problems, but less sensitive to the individual expression values.

3.4.2 The TNoM Test

The TNoM test is nonparametric method[64]. The *Threshold Number of Misclassification* (TNoM) measures how successful we are in separating the two groups of samples by a simple threshold over the expression values. In other words, we search for threshold value of the gene's expression that will distinguish the two groups, A and B, and not for difference between means. A gene is scored by the number of misclassifications made by the best threshold that we can find for it. If the expression value of the gene allows us to separate the groups perfectly, the gene has a TNoM score of zero. On the other hand, if the two groups are interspersed, the gene has a score that may be close to the size of the smallest group of samples (which is the maximum possible number of misclassifications).

More formally, in order to calculate the minimum number of misclassifications, we first define a vector g, such as:

Equation 1

$$g(x:j,t,d) = \begin{cases} d & x[j] > t \\ -d & x[j] < t \end{cases}$$

where *x* is an expression profile , *j* is an index of a gene, *x[j]* is the expression value of the *j*'th gene in the vector *x*, *t* is a threshold corresponding to gene, and $d \in \{+1, -1\}$ is a direction parameter.

The number of errors is defined as:

$$Err(d, t \mid g, l) = \sum_i 1\{l_i \neq \text{sign}(d \cdot (x_i[g] - t))\},$$

Equation 2

where $x_i[j]$ is the expression value of gene *g* in the *i*'th sample and l_i is the label of the *i*'th sample.

The TNoM score of a gene is:

$$TNoM(g, l) = \min_{d,t} Err(d, t \mid g, l)$$

Equation 3

meaning, the minimum number of errors for different values of *t*.

In order to ask whether genes with low TNoM scores are indeed indicative of the classification of expression, we are calculating their P-values, as describe in Ben-Dor et al[64].

3.4.3 FDR

In DNA microarray experiments the number of supervised tests preformed is in the order of thousands. Therefore, the multiplicity problem should be taken into consideration. In order to address contamination with false positive genes associated with multiple comparisons we use the method of Benjamini and Hochberg[66] that bounds the average False Discovery Rate (FDR); namely, the fraction of false positives among the list of differentiating genes. See [66] for further information regarding FDR.

3.4.4 POC

In order to analyze the promoter regions of a cluster of genes and to identify the main common transcriptional regulators of these genes, we used a recently developed program called POC, which searches for significant presence of known TFs' binding-sites in a given cluster's promoters. See [55] for further information regarding POC.

4 Summary

In this work we analyzed recently published gene expression data of different subtypes of childhood ALL. We applied an unsupervised approach, using the SPC and CTWC clustering methods in order to search for a set of genes, whose expression profile separates the samples into two unanticipated distinct groups. This class discovery analysis is of great importance in leukemia and it may contribute to the accurate assignment of individual patients to specific leukemia subtypes. Previous discovery of novel subtypes led to more efficient targeted drug design. Of great interest is the possible recognition of a class of patients, who all share a certain causative agent in the pathogenesis of leukemia.

We observed two unanticipated partitions of the ALL patients to novel subgroups, induced by two different clusters of genes. We examined these two clusters separately:

The first cluster contains many genes that are relevant to the immune response, and in particular immune response to viruses. Most of these are directly related to the interferon-JAK/STAT1 pathway. It has long been suspected that common childhood infections contribute to the etiology of childhood leukemia, in particular ALL. Notably, the interferon genes presented high expression values in the hyperdiploid leukemia variant that occurs in early childhood, when viral infection is most likely to occur. Our findings were confirmed on additional published datasets. A clear signature of overexpression of the interferon genes was found for two other cases: breast cancer and ovary cancer. These observations strengthen previous epidemiological and molecular studies, which suggested a role for infection in the etiology of these cancers. A search for infectious agents, mainly leukemias and breast cancer cases where high expression of the interferon genes are identified, may lead to the identification of the putative viral pathogen. The implications of the finding of such a causative agent on diagnosis, therapy and prevention of childhood leukemia, breast cancer and other malignancies are clear.

The second cluster of genes induced interesting separations in several datasets. The common pathway of these genes is still vague, but by performing a preliminary sequence analysis of the promoter regions of these genes, we revealed potential common transcription factors. These TFs are known to be associated with leukemia, but at the moment we lack essential information, such as additional gene annotations or supplementary clinical data for the samples to indicate the relevant biological process. Hopefully in the near future, using supplemental data and further analysis of the promoter regions, the yet unrecognized pathway, whose existence we demonstrated, will be elucidated and completed and its role in the pathogenesis of leukemia will become clear.

In order to discover new classes, we used a method which combined a clustering algorithm with other statistical tests, and involved the analyses of several different datasets which were produced by several different research groups; among them are 4 datasets of leukemia patients and 8 datasets of other types of cancer. The search for the characteristic gene set was performed in a totally unprejudiced "blind" way, regarding both the separating gene set and the resulting partition of the samples.

Our findings demonstrate the importance of class discovery using unsupervised methods. While 'traditional' supervised methods use prior knowledge about the data and do not allow extracting hidden information underlying in the data, our method, which combined a clustering algorithm and other statistical tests, was able to discover unexpected partitions in the data.

References

1. Rowley JD. The critical role of chromosome translocations in human leukemias. *Annu Rev Genet* 1998;**32:**495-519.
2. http://www.mskcc.org/mskcc/html/6030.cfm. Memorial Sloan-Kerrering Cancer Center,Leukemia, 2003.
3. Greaves MF, Wiemel J. Origins of chromosome translocations in childhood leukaemia. *Nat Rev Cancer* 2003;**3**(9)**:**639-49.
4. Wiemels JL, Greaves M. Structure and possible mechanisms of TEL-AML1 gene fusions in childhood acute lymphoblastic leukemia. *Cancer Res* 1999;**59**(16)**:**4075-82.
5. Greaves M. Childhood leukaemia. *Bmj* 2002;**324**(7332)**:**283-7.
6. Downing JR, Shannon KM. Acute leukemia: a pediatric perspective. *Cancer Cell* 2002;**2**(6)**:**437-45.
7. 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)**:**133-43.
8. Ford AM, Ridge SA, Cabrera ME, et al. In utero rearrangements in the trithorax-related oncogene in infant leukaemias. *Nature* 1993;**363**(6427)**:**358-60.
9. Wiemels JL, Ford AM, Van Wering ER, Postma A, Greaves M. Protracted and variable latency of acute lymphoblastic leukemia after TEL-AML1 gene fusion in utero. *Blood* 1999;**94**(3)**:**1057-62.
10. Buckley JD, Buckley CM, Breslow NE, Draper GJ, Roberson PK, Mack TM. Concordance for childhood cancer in twins. *Med Pediatr Oncol* 1996;**26**(4)**:**223-9.
11. Mori H, Colman SM, Xiao Z, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci U S A* 2002;**99**(12)**:**8242-7.
12. Alexander FE, Greaves MF. Ionising radiation and leukaemia potential risks: review based on the workshop held during the 10th Symposium on Molecular Biology of Hematopoiesis and Treatment of Leukemia and Lymphomas at Hamburg, Germany on 5 July 1997. *Leukemia* 1998;**12**(8)**:**1319-23.
13. Henle G, Henle W. Immunofluorescence in cells derived from Burkitt's lymphoma. *J Bacteriol* 1966;**91**(3)**:**1248-56.
14. Poiesz BJ, Ruscetti FW, Gazdar AF, Bunn PA, Minna JD, Gallo RC. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc Natl Acad Sci U S A* 1980;**77**(12)**:**7415-9.

15. Mele A, Pulsoni A, Bianco E, et al. Hepatitis C virus and B-cell non-Hodgkin lymphomas: an Italian multicenter case-control study. *Blood* 2003;**102**(3)**:**996-9.
16. Peek RM, Jr., Blaser MJ. Helicobacter pylori and gastrointestinal tract adenocarcinomas. *Nat Rev Cancer* 2002;**2**(1)**:**28-37.
17. Kinlen LJ. Epidemiological evidence for an infective basis in childhood leukaemia. *Br J Cancer* 1995;**71**(1)**:**1-5.
18. Kinlen LJ, Balkwill A. Infective cause of childhood leukaemia and wartime population mixing in Orkney and Shetland, UK. *Lancet* 2001;**357**(9259)**:**858.
19. Koushik A, King WD, McLaughlin JR. An ecologic study of childhood leukemia and population mixing in Ontario, Canada. *Cancer Causes Control* 2001;**12**(6)**:**483-90.
20. Greaves MF. Speculations on the cause of childhood acute lymphoblastic leukemia. *Leukemia* 1988;**2**(2)**:**120-5.
21. Greaves MF, Alexander FE. An infectious etiology for common acute lymphoblastic leukemia in childhood? *Leukemia* 1993;**7**(3)**:**349-60.
22. Greaves MF. Aetiology of acute leukaemia. *Lancet* 1997;**349**(9048)**:**344-9.
23. Dworsky M, Yow M, Stagno S, Pass RF, Alford C. Cytomegalovirus infection of breast milk and transmission in infancy. *Pediatrics* 1983;**72**(3)**:**295-9.
24. MacKenzie J, Perry J, Ford AM, Jarrett RF, Greaves M. JC and BK virus sequences are not detectable in leukaemic samples from children with common acute lymphoblastic leukaemia. *Br J Cancer* 1999;**81**(5)**:**898-9.
25. MacKenzie J, Gallagher A, Clayton RA, et al. Screening for herpesvirus genomes in common acute lymphoblastic leukemia. *Leukemia* 2001;**15**(3)**:**415-21.
26. Tabori U, Burstein Y, Dvir R, Rechavi G, Toren A. The clinical and biological dilemma of preleukemia presenting as aplastic anemia with chromosomal translocation t(4;11). *Leukemia* 2001;**15**(5)**:**866-7.
27. Hasle H, Heim S, Schroeder H, Schmiegelow K, Ostergaard E, Kerndrup G. Transient pancytopenia preceding acute lymphoblastic leukemia (pre-ALL). *Leukemia* 1995;**9**(4)**:**605-8.
28. Silverman LB, Gelber RD, Dalton VK, et al. Improved outcome for children with acute lymphoblastic leukemia: results of Dana-Farber Consortium Protocol 91-01. *Blood* 2001;**97**(5)**:**1211-8.
29. Pui CH, Evans WE. Acute lymphoblastic leukemia. *N Engl J Med* 1998;**339**(9)**:**605-15.
30. Fletcher L. Approval heralds new generation of kinase inhibitors? *Nat Biotechnol* 2001;**19**(7)**:**599-600.
31. Galloway DA. Papillomavirus vaccines in clinical trials. *Lancet Infect Dis* 2003;**3**(8)**:**469-75.
32. Devaraj K, Gillison ML, Wu TC. Development of HPV vaccines for HPV-associated head and neck squamous cell carcinoma. *Crit Rev Oral Biol Med* 2003;**14**(5)**:**345-62.

33. Macsween KF, Crawford DH. Epstein-Barr virus-recent advances. *Lancet Infect Dis* 2003;**3**(3)**:**131-40.

34. 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.

35. 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.

36. Rozovskaia T, Ravid-Amir O, Tillib S, et al. Expression profiles of acute lymphoblastic and myeloblastic leukemias with ALL-1 rearrangements. *Proc Natl Acad Sci U S A* 2003.

37. Veldman T, Vignon C, Schrock E, Rowley JD, Ried T. Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping. *Nat Genet* 1997;**15**(4)**:**406-10.

38. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* 2001;**98**(24)**:**13790-5.

39. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 2001;**98**(26)**:**15149-54.

40. Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002;**8**(1)**:**68-74.

41. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002;**1**(2)**:**203-9.

42. Staunton JE, Slonim DK, Coller HA, et al. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A* 2001;**98**(19)**:**10787-92.

43. Welsh JB, Zarrinkar PP, Sapinoso LM, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc Natl Acad Sci U S A* 2001;**98**(3)**:**1176-81.

44. Welsh JB, Sapinoso LM, Su AI, et al. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* 2001;**61**(16)**:**5974-8.

45. 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.

46. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. *Nat Genet* 2002;**31**(4)**:**370-7.

47. Zhang H, Yang B, Pomerantz RJ, Zhang C, Arunachalam SC, Gao L. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature* 2003;**424**(6944)**:**94-8.

48. Mangeat B, Turelli P, Caron G, Friedli M, Perrin L, Trono D. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* 2003;**424**(6944)**:**99-103.

49. Ness RB, Goodman MT, Shen C, Brunham RC. Serologic evidence of past infection with Chlamydia trachomatis, in relation to ovarian cancer. *J Infect Dis* 2003;**187**(7)**:**1147-52.

50. Kasai K, Sato Y, Kameya T, et al. Incidence of latent infection of Epstein-Barr virus in lung cancers--an analysis of EBER1 expression in lung cancers by in situ hybridization. *J Pathol* 1994;**174**(4)**:**257-65.

51. Ford CE, Tran D, Deng Y, Ta VT, Rawlinson WD, Lawson JS. Mouse mammary tumor virus-like gene sequences in breast tumors of Australian and Vietnamese women. *Clin Cancer Res* 2003;**9**(3)**:**1118-20.

52. Keydar I, Ohno T, Nayak R, et al. Properties of retrovirus-like particles produced by a human breast carcinoma cell line: immunological relationship with mouse mammary tumor virus proteins. *Proc Natl Acad Sci U S A* 1984;**81**(13)**:**4188-92.

53. Wang Y, Holland JF, Bleiweiss IJ, et al. Detection of mammary tumor virus env gene-like sequences in human breast cancer. *Cancer Res* 1995;**55**(22)**:**5173-9.

54. Benekli M, Baer MR, Baumann H, Wetzler M. Signal transducer and activator of transcription proteins in leukemias. *Blood* 2003;**101**(8)**:**2940-54.

55. Tabach Y. M.Sc Thesis, Weizmann Institute of Science. 2004.

56. Kumar AR, Hudson WA, Chen W, Nishiuchi R, Yao Q, Kersey JH. Hoxa9 influences the phenotype but not the incidence of Mll-AF9 fusion gene leukemia. *Blood* 2003.

57. Kroon E, Krosl J, Thorsteinsdottir U, Baban S, Buchberg AM, Sauvageau G. Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. *Embo J* 1998;**17**(13)**:**3714-25.

58. Joh T, Hosokawa Y, Suzuki R, Takahashi T, Seto M. Establishment of an inducible expression system of chimeric MLL-LTG9 protein and inhibition of Hox a7, Hox b7 and Hox c9 expression by MLL-LTG9 in 32Dcl3 cells. *Oncogene* 1999;**18**(4)**:**1125-30.

59. Rozovskaia T, Feinstein E, Mor O, et al. Upregulation of Meis1 and HoxA9 in acute lymphocytic leukemias with the t(4 : 11) abnormality. *Oncogene* 2001;**20**(7)**:**874-8.

60. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;**17**(6)**:**520-5.

61. Blatt M, Wiseman S, Domany E. Superparamagnetic clustering of data. *Physical Review Letters* 1996;**76**(18)**:**3251-3254.

62. Blatt M, Wiseman S, Domany E. Data Clustering Using a Model Granular Magnet. *Neural Computation* 1997;**9**(8)**:**1805-1842.

63. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A* 2000;**97**(22)**:**12079-84.

64. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. *J Comput Biol* 2000;**7**(3-4)**:**559-83.

65. Kaminski N, Friedman N. Practical approaches to analyzing results of microarray experiments. *Am J Respir Cell Mol Biol* 2002;**27**(2)**:**125-32.

66. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy Stat Soc* 1995;**57**(Ser B):289-300.

Appendix

Table 1| A list of 79 TEL-AML1 samples from the Yeoh et al.[7] dataset. The samples from the novel sub-group that overexpressed the genes of cluster no. 1 are in **bold face**.

No.	*Clinical label*	*Label ID*
1	**TEL-AML1 CCR T13B**	**TEL-AML1-C52**
2	**TEL-AML1 CCR T13A**	**TEL-AML1-C24**
3	**TEL-AML1 CCR T13B**	**TEL-AML1-C38**
4	**TEL-AML1 CCR T13A**	**TEL-AML1-C1**
5	**TEL-AML1 CCR T13A**	**TEL-AML1-C16**
6	**TEL-AML1 CCR T13B**	**TEL-AML1-C51**
7	**TEL-AML1 CCR T13A**	**TEL-AML1-C15**
8	**TEL-AML1 CCR T13B**	**TEL-AML1-C56**
9	TEL-AML1 CCR T13B	TEL-AML1-C39
10	TEL-AML1 CCR T13B	TEL-AML1-C31
11	TEL-AML1 CCR T13A	TEL-AML1-C28
12	TEL-AML1 T13A	TEL-AML1-#2
13	TEL-AML1 CCR T13B	TEL-AML1-C48
14	TEL-AML1 CCR T13A	TEL-AML1-C5
15	TEL-AML1 CCR T13B	TEL-AML1-C45
16	TEL-AML1 Hematologc Relapse T13B	TEL-AML1-R3
17	TEL-AML1 CCR T13B	TEL-AML1-C35
18	TEL-AML1 CCR T13A	TEL-AML1-C17
19	TEL-AML1 CCR T13B	TEL-AML1-C46
20	TEL-AML1 CCR T13A	TEL-AML1-C6
21	TEL-AML1 CCR T13B	TEL-AML1-C53
22	TEL-AML1 T15	TEL-AML1-#11
23	TEL-AML1 CCR T13A	TEL-AML1-C20
24	TEL-AML1 CCR T13B	TEL-AML1-C33
25	TEL-AML1 T15	TEL-AML1-#7
26	TEL-AML1 second AML T13A	TEL-AML1-2M#3
27	TEL-AML1 T15	TEL-AML1-#6
28	TEL-AML1 CCR T13B	TEL-AML1-C57
29	TEL-AML1 CCR T13B	TEL-AML1-C55
30	TEL-AML1 CCR T13A	TEL-AML1-C11
31	TEL-AML1 CCR T13A	TEL-AML1-C26
32	TEL-AML1 CCR T13B	TEL-AML1-C54
33	TEL-AML1 T15	TEL-AML1-#9
34	TEL-AML1 T15	TEL-AML1-#12
35	TEL-AML1 CCR T13B	TEL-AML1-C44
36	TEL-AML1 CCR T13A	TEL-AML1-C2
37	TEL-AML1 CCR T13A	TEL-AML1-C23
38	TEL-AML1 CCR T13B	TEL-AML1-C42
39	TEL-AML1 CCR T13A	TEL-AML1-C21
40	TEL-AML1 CCR T13A	TEL-AML1-C3
41	TEL-AML1 T15	TEL-AML1-#8
42	TEL-AML1 second AML T13B	TEL-AML1-2M#4
43	TEL-AML1 CCR T13A	TEL-AML1-C27
44	TEL-AML1 CCR T13A	TEL-AML1-C19
45	TEL-AML1 T15	TEL-AML1-#5
46	TEL-AML1 CCR T13B	TEL-AML1-C37
47	TEL-AML1 CCR T13B	TEL-AML1-C32
48	TEL-AML1 T15	TEL-AML1-#10
49	TEL-AML1 second AML T13A	TEL-AML1-2M#1
50	TEL-AML1 CCR T13A	TEL-AML1-C13
51	TEL-AML1 CCR T13A	TEL-AML1-C25
52	TEL-AML1 Hematologc Relapse T13A	TEL-AML1-R2
53	TEL-AML1 T15	TEL-AML1-#13
54	TEL-AML1 CCR T13A	TEL-AML1-C22
55	TEL-AML1 CCR T13B	TEL-AML1-C34
56	TEL-AML1 CCR T13B	TEL-AML1-C30

57	TEL-AML1 CCR T13B	TEL-AML1-C29
58	TEL-AML1 CCR T13B	TEL-AML1-C47
59	TEL-AML1 CCR T13B	TEL-AML1-C40
60	TEL-AML1 CCR T13A	TEL-AML1-C7
61	TEL-AML1 second AML T13B	TEL-AML1-2M#5
62	TEL-AML1 second AML T13A	TEL-AML1-2M#2
63	TEL-AML1 CCR T13A	TEL-AML1-C12
64	TEL-AML1 Hematologc Relapse T13A	TEL-AML1-R1
65	TEL-AML1 CCR T13B	TEL-AML1-C43
66	TEL-AML1 CCR T13A	TEL-AML1-C18
67	TEL-AML1 T13A	TEL-AML1-#3
68	TEL-AML1 CCR T13B	TEL-AML1-C36
69	TEL-AML1 CCR T13B	TEL-AML1-C50
70	TEL-AML1 T13B	TEL-AML1-#1
71	TEL-AML1 T15	TEL-AML1-#14
72	TEL-AML1 CCR T13A	TEL-AML1-C4
73	TEL-AML1 CCR T13B	TEL-AML1-C41
74	TEL-AML1 CCR T13A	TEL-AML1-C14
75	TEL-AML1 CCR T13A	TEL-AML1-C10
76	TEL-AML1 T13B	TEL-AML1-#4
77	TEL-AML1 CCR T13B	TEL-AML1-C49
78	TEL-AML1 CCR T13A	TEL-AML1-C9
79	TEL-AML1 CCR T13A	TEL-AML1-C8

Subtype Name-C# Dx Sample of patient in CCR (Continuous complete remission)
Subtype Name-R# Dx Sample of patient who developed a hematologic relapse
Subtype Name-# Dx Sample used for subgroup classification only
Subtype Name-2M# Dx Sample of patient who later developed 2nd AML
Subtype Name-N Dx Sample in novel groupas was found by Yeoh et al.
T## - Protocol that patient was treated on

Table 2| Hyperdiploidity samples from the Yeoh et al.[7] dataset. The samples from the novel sub-group that underexpressed the genes of cluster no. 2 are in **bold face**. The samples are ordered by CTWC.

No.	*Samples' clinical labels*	*Samples' ID*
1	Hyperdyp>50 T15	Hyperdip>50-#11
2	Hyperdyp>50 T15	Hyperdip>50-#6
3	Hyperdyp>50 CCR T13B	Hyperdip>50-C43
4	Hyperdyp>50 CCR T13A	Hyperdip>50-C1
5	Hyperdyp>50 CCR T13A	Hyperdip>50-C12
6	Hyperdyp>50 CCR T13B	Hyperdip>50-C34
7	Hyperdyp>50 CCR T13B	Hyperdip>50-C17
8	Hyperdyp>50 T15	Hyperdip>50-#10
9	Hyperdyp>50 CCR T13A	Hyperdip>50-C3
10	Hyperdyp>50 Hematologc Relapse T13B	Hyperdip>50-R5
11	Hyperdyp>50 second AML T13A	Hyperdip>50-2M#1
12	Hyperdyp>50 CCR T13A	Hyperdip>50-C4
13	Hyperdyp>50 CCR T13A	Hyperdip>50-C6
14	Hyperdyp>50 CCR T13B	Hyperdip>50-C23
15	Hyperdyp>50 CCR T13B	Hyperdip>50-C15
16	Hyperdyp>50 T13B	Hyperdip>50-#2
17	Hyperdyp>50	Hyperdip>50-#4
18	Hyperdyp>50 T15	Hyperdip>50-#14
19	Hyperdyp>50 CCR T13A	Hyperdip>50-C13
20	Hyperdyp>50 CCR T13B	Hyperdip>50-C42
21	Hyperdyp>50 CCR T13B	Hyperdip>50-C21
22	Hyperdyp>50 T15	Hyperdip>50-#12
23	Hyperdyp>50 CCR T13B	Hyperdip>50-C25
24	Hyperdyp>50 CCR T13B	Hyperdip>50-C20
25	Hyperdyp>50	Hyperdip>50-#3
26	Hyperdyp>50 CCR T13B	Hyperdip>50-C40
27	Hyperdyp>50 T15	Hyperdip>50-#9
28	Hyperdyp>50 T12	Hyperdip>50-#5
29	Hyperdyp>50 CCR T13A	Hyperdip>50-C5
30	Hyperdyp>50 T15	Hyperdip>50-#7
31	Hyperdyp>50 CCR T13B	Hyperdip>50-C19
32	Hyperdyp>50 CCR T13B	Hyperdip>50-C16

33	Hyperdyp>50 CCR T13B	Hyperdip>50-C32
34	Hyperdyp>50 CCR T13B	Hyperdip>50-C33
35	Hyperdyp>50 CCR T13A	Hyperdip>50-C8
36	Hyperdyp>50 T15	Hyperdip>50-#13
37	Hyperdyp>50 CCR T13B	Hyperdip>50-C35
38	Hyperdyp>50 CCR T13B	Hyperdip>50-C24
39	Hyperdyp>50 CCR T13A	Hyperdip>50-C9
40	Hyperdyp>50 CCR T13A	Hyperdip>50-C7
41	Hyperdyp>50 CCR Novel Group T13B	Hyperdip>50-C27-N
42	Hyperdyp>50 T13A	Hyperdip>50-#1
43	Hyperdyp>50 second AML T12	Hyperdip>50-2M#3
44	Hyperdyp>50 T15	Hyperdip>50-#8
45	Hyperdyp>50 CCR T13A	Hyperdip>50-C11
46	Hyperdyp>50 CCR T13B	Hyperdip>50-C37
47	Hyperdyp>50 CCR T13B	Hyperdip>50-C30
48	Hyperdyp>50 CCR T13B	Hyperdip>50-C41
49	Hyperdyp>50 CCR T13B	Hyperdip>50-C31
50	Hyperdyp>50 CCR T13B	Hyperdip>50-C28
51	**Hyperdyp>50 CCR T13B**	**Hyperdip>50-C29**
52	**Hyperdyp>50 CCR T13B**	**Hyperdip>50-C39**
53	**Hyperdyp>50 Hematologc Relapse T13B**	**Hyperdip>50-R4**
54	**Hyperdyp>50 Hematologc Relapse T13A**	**Hyperdip>50-R2**
55	**Hyperdyp>50 CCR T13B**	**Hyperdip>50-C22**
56	**Hyperdyp>50 CCR T13B**	**Hyperdip>50-C26**
57	**Hyperdyp>50 second AML T13B**	**Hyperdip>50-2M#2**
58	**Hyperdyp>50 CCR T13A**	**Hyperdip>50-C14**
59	**Hyperdyp>50 CCR T13A**	**Hyperdip>50-C10**
60	**Hyperdyp>50 CCR T13B**	**Hyperdip>50-C18**
61	**Hyperdyp>50 CCR T13B**	**Hyperdip>50-C36**
62	**Hyperdyp>50 Hematologc Relapse T13A**	**Hyperdip>50-R1**
63	**Hyperdyp>50 CCR T13B**	**Hyperdip>50-C38**
64	**Hyperdyp>50 CCR T13A**	**Hyperdip>50-C2**
65	**Hyperdyp>50 Hematologc Relapse T13A**	**Hyperdip>50-R3**

Subtype Name-C# Dx Sample of patient in CCR (Continuous complete remission)
Subtype Name-R# Dx Sample of patient who developed a hematologic relapse
Subtype Name-# Dx Sample used for subgroup classification only
Subtype Name-2M# Dx Sample of patient who later developed 2nd AML
Subtype Name-N Dx Sample in novel groupas was found by Yeoh et al.
T## - Protocol that patient was treated on

Table 3 | The distribution of samples within the two novel sub-groups. The sub-groups were obtained by the analysis of the Yeoh et al.[7] dataset, using the genes of cluster no. 2.

Subtype name	***Number of samples***	***Number in subgroup***
Hyperdip>50	65	13
TEL-AML1	79	6
Pseudodip	29	2
Normal	19	1
Hyperdip 47-50	23	4
T-ALL	45	3
BCR-ABL	16	3
MLL	21	3
Hypodip	11	1
E2A-PBX1	27	1
Total:	**335**	**37**

Table 4| The samples and their labels of Rozovskia et al.[36] dataset. The samples that underexpressed the genes of cluster no. 2 are in **bold face**. The samples within the subgroups are reordered according to their labels.

No.	Main Label	patient/cell line	Label ID
1	CD10-ALL	Patient	12ht
2	CD10-ALL	Patient	14ht
3	CD10-ALL	Patient	16ht
4	T-cell ALL	Patient	03ht
5	M5AML	Patient	48.1ht
6	M5AML	Patient	48.2ht
7	M5aAML	Patient	51ht
8	M5aAML	Patient	52ht
9	t(6;11)AML	ML2-cell line	44ht
10	t(9;11)M5AML	Patient	35ht
11	t(9;11)M5AML	Patient	36ht
12	t(9;11)M5AML	Patient	37ht
13	t(9;11)M5AML	Patient	38ht
14	t(9;11)AML	TPH1-cell line	38.1ht
15	t(9;11)AML	MONO6-cell line	38.2ht
16	t(9;11)AML	PER377-cell line	38.4ht
17	t(4;11)ALL	Patient	13ht
18	t(4;11)ALL	Patient	18ht
19	t(4;11)ALL	Patient	19ht
20	t(4;11)ALL	Patient	20ht
21	t(4;11)ALL	Patient	22ht
22	t(4;11)ALL	Patient	23ht
23	t(4;11)ALL	Patient	24ht
24	t(4;11)ALL	Patient	25ht
25	t(4;11)ALL	Patient	26ht
26	t(4;11)ALL	B1-cell line	27.1ht
27	t(4;11)ALL	RS(4;11)-cell line	27.2ht
28	Nor		33.1ht
29	Nor		33ht
30	Nor		33.2ht
31	Nor		28ht
32	Nor		29ht
33	Nor		30ht
34	Nor		32ht
35	Nor		31ht
36	**T-cell ALL**	**Patient**	**01ht**
37	**T-cell ALL**	**Patient**	**02ht**
38	**CD10+ALL**	**Patient**	**05ht**
39	**CD10+ALL**	**Patient**	**06ht**
40	**CD10+ALL**	**Patient**	**07ht**
41	**Ph+ALL**	**Patient**	**09ht**
42	**Ph+ALL**	**Patient**	**10ht**
43	**Ph+ALL**	**Patient**	**11ht**
44	**CD10-ALL**	**Patient**	**15ht**
45	**M4AML**	**Patient**	**45ht**
46	**M4AML**	**Patient**	**46ht**
47	**M4AML**	**Patient**	**47ht**
48	**M5AML**	**Patient**	**48ht**
49	**M5aAML**	**Patient**	**49ht**
50	**M5aAML**	**Patient**	**50ht**
51	**MLLdup.M4AML**	**Patient**	**39ht**
52	**MLLdup.M5bAML**	**Patient**	**40ht**
53	**MLLdup.M5aAML**	**Patient**	**41ht**
54	**t(10;11)M5AML**	**Patient**	**42ht**
55	**t(11;19)AML**	**Patient**	**43ht**
56	**t(4;11)ALL**	**Patient**	**17ht**
57	**t(4;11)ALL**	**Patient**	**21ht**
58	**t(4;11)ALL**	**Patient**	**27ht**
59	**t(9;11)M5AML**	**Patient**	**34ht**
60	**t(9;11)AML**	**MOL13-cell line**	**38.3ht**

גילוי תתי-סיווגים בלוקימיה לימפובְּלַסטית חריפה
באמצעות חקר ביטוי גנטי

אורי עינב

תזה לשם קבלת התואר מוסמך במדעים מוגש למועצה המדעית של

מכון ויצמן למדע

בהדרכת

פרופסור איתן דומאני

דצמבר 2003