Significance of Murine Retroviral Mutagenesis for Identification of Disease Genes in Human Acute Myeloid Leukemia

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

Department of Hematology, Erasmus University Medical Center, Rotterdam, the Netherlands

Abstract

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

Introduction

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

Materials and Methods

GEP data from AML patients. Data from Affymetrix HGU133A GeneChip analysis in 285 adult AML patients are available (10). Patients were categorized as favorable and unfavorable risk based on cytogenetic variables. The favorable risk group comprised cases with t(8;21), inv(16), and t(15;17) without additional unfavorable cytogenetic abnormalities. The unfavorable group comprised cases with complex karyotype abnormalities, −5 or 5q−, −7 or 7q−, t(6;9), t(9;22), and 11q23 abnormalities but no unfavorable risk abnormalities. GEP data from 130 childhood AML samples are available (9). Data were normalized by global scaling (Affymetrix Microarray Suite version 5.0; MAS5.0) with target average intensity values of 100 for the adult AML and 500 for the pediatric AML data set, respectively. Expression levels from each probe set in every sample were calculated relative to the geometric mean and logarithmically transformed (base 2) to ascribe equal weight to gene expression levels with equal relative distance to the geometric mean. Significance analysis of microarrays (SAM; ref. 11) was used to identify genes contributing to the unsupervised clustering of patients in the adult and pediatric AML groups. In the adult AML data set, 16 classes resulting from unsupervised clustering (10) were evaluated. In the juvenile AML data set, five classes defined according to cytogenetic aberrations (9) were analyzed. This analysis was done for all probe sets represented on the HGU133A GeneChip (n = 22,283). Patients from a specific class were tested compared with all remaining samples using an 8 test and sample class permutations to assess statistical significance. Probe sets were considered differential when fold change values exceeded 1.5 or were <0.67, scores were >4 or less than −4, and q values were <0.05, where false discovery rate was <0.05.

Note: S.J. Erkeland and R.G.W. Verhaak contributed equally to this work.

Requests for reprints: Ivo P. Touw, Department of Hematology, Erasmus University Medical Center, Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, the Netherlands. Phone: 31-10-408-8738, Fax: 31-10-408-9470; E-mail: touwi@erasmusmc.nl.

Significance of difference in number of differentially expressed probe sets. To calculate the significance of difference in the number of differentially expressed probe sets in two groups (i.e., VIS representing probe sets versus probe sets not representing a VIS), Pearson’s $\chi^2$ with 1 degree of freedom was calculated using 2 × 2 contingency tables. As some probe sets were differential in multiple clusters, all possibilities on differential expression were taken into account. For instance, 16 SAM values used in the contingency table was 16

**Results**

**VIS Genes Contribute to Clustering of AML by GEP**

**Gr-1.4 VIS genes and adult AML.** To assess the relevance of Gr-1.4 VIS and CIS genes for human AML, we determined their expression in different classes of adult AML patients (9, 10). Based on unsupervised cluster analysis of GEP data, 285 adult AML cases were grouped in 16 subclasses (10). With SAM, specific gene sets were linked to these subclasses, by comparing each subclass with the remaining cases. In total, 5,193 probe sets, representing 3,644 genes, contributed to the signature of the 16 subclasses (Supplementary Table 1a). We calculated that the probability that a randomly selected gene is differentially expressed in one or more subclasses is 0.28 (Table 1) and did Pearson’s $\chi^2$ analysis to test whether VIS and CIS genes have a higher than random probability to be differentially expressed in one of the subclasses. Four gene lists derived from the Gr-1.4-induced leukemia model and represented on the HGU133A GeneChip were analyzed: (I) VIS + CIS genes ($n = 115$, represented by 234 probe sets); (II) CIS genes ($n = 53$, 116 probe sets); (III) direct neighbors of CIS genes ($n = 53$, 81 probe sets); (IV) genes located within a region of 1 Mb of the CIS genes, with a maximum of five genes upstream or downstream ($n = 279$, 468 probe sets; Fig. 1; Supplementary Table 2a-d). The VIS and CIS genes have a significantly increased probability ($0.46; P = 0.001$ and $0.43, P = 0.002$, respectively) to be differentially expressed in subclasses of adult AML compared with unselected genes (I and II in Table 1; genes are listed in Supplementary Table 3a and b). In contrast, no such correlation was found for gene lists III and IV (Table 1).

**Gr-1.4 VIS genes and pediatric AML.** To determine the validity of these results for an independent AML GEP data set, correlation analysis was done on 130 childhood AML samples (9). Patients were grouped in five subclasses [i.e., cases with inv(16), t(15;17), t(8;21)], translocations involving MLL, and cases with megakaryoblastic leukemia (Supplementary Table 1b). In total, 2,736 probe sets, representing 2,093 genes, contributed to the signature of the five subclasses. The probability that a randomly selected gene is differentially expressed in one or more subclasses of the childhood AML data set was 0.16 (Table 1). Similar to adult AML, Gr-1.4 CIS and VIS genes had a significantly increased probability (0.31; $P = 0.0127$ and 0.25, $P = 0.005$, respectively) to be differentially expressed in the distinct patient clusters, whereas again no such correlation was seen with more distantly located genes (Supplementary Table 3e and f).

**BXH2 and AkxD VIS genes and AML.** Candidate leukemia genes identified in two other models, BXH2 and AkxD (Supplementary Table 2e and f) also correlated significantly with the gene sets responsible for clustering of adult (0.62; $P < 0.0001$ and 0.61, $P < 0.0001$, for BXH2 and AkxD CIS/VIS, respectively) and pediatric

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

<table>
<thead>
<tr>
<th></th>
<th>No. unique genes (probe sets)</th>
<th>No. unique SAM genes in adult AML (probe sets)</th>
<th>Probability $^a$</th>
<th>P $^a$</th>
<th>No. unique SAM genes in pediatric AML (probe sets)</th>
<th>Probability</th>
<th>P $^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All genes</td>
<td>12,848 (22,283)</td>
<td>3,644 (5,193) $^b$</td>
<td>0.28</td>
<td>—</td>
<td>2,093 (2,736) $^b$</td>
<td>0.16</td>
<td>—</td>
</tr>
<tr>
<td>(I) Gr-1.4 CIS genes</td>
<td>51 (116)</td>
<td>22 (32)</td>
<td>0.43</td>
<td>0.002</td>
<td>16 (21)</td>
<td>0.31</td>
<td>0.0127</td>
</tr>
<tr>
<td>(II) Gr-1.4 VIS genes</td>
<td>115 (234)</td>
<td>53 (74)</td>
<td>0.46 &lt;0.0001</td>
<td></td>
<td>29 (42)</td>
<td>0.25</td>
<td>0.0050</td>
</tr>
<tr>
<td>(III) 2 adjacent genes</td>
<td>53 (81)</td>
<td>15 (18)</td>
<td>0.28</td>
<td>0.49 (NS)</td>
<td>7 (9)</td>
<td>0.13</td>
<td>0.9361 (NS)</td>
</tr>
<tr>
<td>(IV) 10 adjacent genes</td>
<td>279 (468)</td>
<td>91 (123)</td>
<td>0.33</td>
<td>0.19 (NS)</td>
<td>50 (66)</td>
<td>0.18</td>
<td>0.2071 (NS)</td>
</tr>
</tbody>
</table>

Candidate leukemia genes from other mouse models

<table>
<thead>
<tr>
<th></th>
<th>No. unique genes (probe sets)</th>
<th>No. unique SAM genes (probe sets)</th>
<th>Probability $^*$</th>
<th>P $^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BXH2 CIS/VIS</td>
<td>53 (111)</td>
<td>33 (51)</td>
<td>0.62 &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>AKxD CIS/VIS</td>
<td>119 (232)</td>
<td>72 (104)</td>
<td>0.61 &lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>All CIS/VIS</td>
<td>237 (470)</td>
<td>122 (178)</td>
<td>0.51 &lt;0.0001</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: NS, not significant.

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

$^b$Because some probe sets contribute to multiple classes, the total number of sets used in the $\chi^2$ analysis was 8,739 for adult AML and 2,955 for the pediatric AML cases. For details, see Supplementary Table 1a and b.

**Virus flanking genes in mouse leukemia.** Genes affected by virus integrations in Grafti L.4 (Gr-1.4), BXH2, and AkxD murine leukemia virus (MuLV) models have been previously reported (3, 12).

**Network and principal component analyses.** Ingenuity pathway analysis$^3$ was used in combination with the Ingenuity Pathways Knowledge Base (IPKB). Genes selected from experimental data, called focus genes, are used for the generation of networks with a maximal size of 35 genes/proteins. Focus genes were VIS genes that significantly contributed to the unsupervised clustering of 285 AML cases. Principal component analysis was done using Spotfire Software (Spotfire, Inc., Somerville, MA).

3 http://genome2.ncifcrf.gov/RTCGD.
AML cases (0.40; \( P = 0.0001 \) and 0.36, \( P < 0.0001 \), respectively; Table 1; Supplementary Table 3c-f). The combined data from the three models indicate that genes directly flanking the virus integrations are significantly more differentially expressed than random genes in both adult and pediatric AML subtypes.

No correlation between proviral integration and actively transcribed genes in normal hematopoietic precursors. To investigate whether correlations between murine VIS genes and human AML panel into the Ingenuity application to place them in regulatory networks. From this list (\( n = 125 \)), 110 genes present in the IPKB (focus genes) were used for the generation of networks. Five highly significant networks, associated with cell growth and proliferation, hematopoietic cell development, cell cycle, and gene expression were identified (Table 2; Supplementary Figs. 1-5). Network 1 existed exclusively of focus genes (\( n = 35 \)), suggesting that genes within this network are commonly deregulated in AML. Multiple genes in this network (i.e., IL2RG, STAT5A, STAT5B, IL4R, IRS2) are involved in cytokine signaling. The SOX4 gene encodes a

### Table 2. VIS/SAM genes in regulatory networks

<table>
<thead>
<tr>
<th>Focus genes in network*</th>
<th>Major global functions of network</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network 1</strong></td>
<td></td>
</tr>
<tr>
<td>BTG2, CEBPB, DOK1, DSPI, DUSP10, E2F2, ELF4, EVI1, FOS, FOSS1, HCK, HMGAI, HRAS, IL2RA, IL2RB, IL2RG, IL4R, IRS2, JUNB, LCK, LEF1, LTβ, MADH3, MPL, NFKB2, NFKB1A, PLA, RUNX1, SOX4, STAT5A, STAT5B, TP53, TRA1, ZFHX1B, ZNF145</td>
<td>Tissue morphology (( n = 25 )) Cellular growth and proliferation (( n = 28 )) Cellular development (( n = 27 ))</td>
</tr>
<tr>
<td><strong>Network 2</strong></td>
<td></td>
</tr>
<tr>
<td>CALD1, CCND2, CCND3, CTNNAI, ETS1, FLJ1, HES1, MYR, MYC, MYCN, NFATC1, NOTCH1, NOTCH2, PAX5, PIM1, PRDX1, PRDX2</td>
<td>Cellular development (( n = 13 )) Hematologic system development and function (( n = 10 )) Cancer (( n = 12 ))</td>
</tr>
<tr>
<td><strong>Network 3</strong></td>
<td></td>
</tr>
<tr>
<td>BCL11A, CAPG, CCL4, CCL5, IFNGR2, IL6ST, INPP5A, KIT, MAP4K2, PTP4A3, PTPRE, PXN, SOC52, SWAP70</td>
<td>Hematologic system development and function (( n = 22 )) Cell death (( n = 23 )) Immune response (( n = 21 ))</td>
</tr>
<tr>
<td><strong>Network 4</strong></td>
<td></td>
</tr>
<tr>
<td>C3AR1, EPS15, HHB1, IFI30, MEF2C, MEF2D, NCOA1, NP, RXRA, ST13, TIE, ZFP36</td>
<td>Gene expression (( n = 17 )) Cellular development (( n = 12 )) Cancer (( n = 13 ))</td>
</tr>
<tr>
<td><strong>Network 5</strong></td>
<td></td>
</tr>
<tr>
<td>BCOR, CCND3, E2F2, GFI1, HOXA9, LMO2, MEIS1, TWIST1</td>
<td>Cancer (( n = 21 )) Gene expression (( n = 22 )) Cellular growth and proliferation (( n = 22 ))</td>
</tr>
</tbody>
</table>

*Complete networks are shown in Supplementary Figs. 1 to 5.*
transcriptional regulator implicated in the pathogenesis of neuronal tumors and lymphoma (13, 14), ZNF145, which is involved in t(11;17) in acute promyelocytic leukemia, encodes a transcriptional repressor also known as promyelocytic leukemia zinc finger (PLZF) that has recently been implicated as a regulator of stem cell renewal (15, 16). We also asked whether networks might be differentially affected in prognostic subgroups of AML. To this end, we applied principal component analysis, by which AML samples are clustered in a three-dimensional space based on expression correlations of genes of each of the separate networks. Thus far, only network 5 clearly discriminated between AML patients with favorable and unfavorable cytogenetic risk indication (Fig. 2). SAM analysis indicated that this distinction is predominantly based on differential expression of HOXA9, MEIS1, and CCND3, which are up-regulated in the unfavorable group, and BCOR and GFI1, which are down-regulated in the unfavorable group (Supplementary Table 6a and b).

Discussion

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

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

Consistent with previous molecular and cytogenetic studies, the networks affected in AML mainly comprise signaling molecules and transcription regulators involved in growth factor–controlled cell proliferation and survival and the transcriptional control of myeloid differentiation (19). However, Gr-1.4 VIS genes deregulated in AML also include genes involved in other mechanisms (Table 2; Supplementary Table 3a and b). For instance, TXNIP and PRDX2 act in cellular responses to oxidative stress, whereas CTNNA1 has been implicated in cell differentiation. CTNNA1 is a candidate tumor suppressor gene located at chromosome 5q3.1 in a region that is frequently deleted in myelodysplasia and AML (20).

An important implication of this work is that disease genes and nonpathogenic genes (e.g., related to differentiation status of the cells) may be distinguished in clinical AML data sets. With the VIS gene lists in the various mouse leukemia models not yet saturated and the possibilities of GEP of AML still growing, the power of this strategy may increase. This may allow further refinement of currently identified and presumably disclose additional pathogenic networks underlying AML. Such information would be useful for further refinement of diagnosis and for identification of key targets for therapeutic intervention.

Acknowledgments

Received 8/16/2005; revised 10/26/2005; accepted 11/16/2005.

Grant support: Dutch Cancer Society "Koningin Wilhelmina Fonds."

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References


Significance of Murine Retroviral Mutagenesis for Identification of Disease Genes in Human Acute Myeloid Leukemia


**Updated version**
Access the most recent version of this article at:
http://cancerres.aacrjournals.org/content/66/2/622

**Supplementary Material**
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2006/01/24/66.2.622.DC1

**Cited articles**
This article cites 20 articles, 7 of which you can access for free at:
http://cancerres.aacrjournals.org/content/66/2/622.full#ref-list-1

**Citing articles**
This article has been cited by 8 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/66/2/622.full#related-urls

**E-mail alerts**
Sign up to receive free email-alerts related to this article or journal.

**Reprints and Subscriptions**
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

**Permissions**
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.