Digital gene expression profiling of primary acute lymphoblastic leukemia cells

J Nordlund 1, A Kialainen 1,9, O Karlberg 1, EC Berglund 1, H Göransson-Kultima 2, M Sonderkaer 3, KL Nielsen 3, MG Gustafsson 2, M Behrendtz 4, E Forestier 5, M Perkkio 6, S Söderhäll 7, G Lönnroth 8 and A-C Syvänen 1

We determined the genome-wide digital gene expression (DGE) profiles of primary acute lymphoblastic leukemia (ALL) cells from 21 patients taking advantage of ’second-generation’ sequencing technology. Patients included in this study represent four cytogenetically distinct subtypes of B-cell precursor (BCP) ALL and T-cell lineage ALL (T-ALL). The robustness of DGE combined with supervised classification by nearest shrunken centroids (NSC) was validated experimentally and by comparison with published expression data for large sets of ALL samples. Genes that were differentially expressed between BCP ALL subtypes were enriched to distinct signalling pathways with dic(9;20) enriched to TP53 signaling, t(9;22) to interferon signaling, as well as high hyperdiploidy and t(12;21) to apoptosis signaling. We also observed antisense tags expressed from the non-coding strand of ~50% of annotated genes, many of which were expressed in a subtype-specific pattern. Antisense tags from 17 gene regions unambiguously discriminated between the BCP ALL and T-ALL subtypes, and antisense tags from 76 gene regions discriminated between the 4 BCP subtypes. We observed a significant overlap of gene regions with alternative polyadenylation and antisense transcription (P < 1 × 10−15). Our study using DGE profiling provided new insights into the RNA expression patterns in ALL cells.

Leukemia (2012) 26, 1218–1227; doi:10.1038/leu.2011.358; published online 16 December 2011

Keywords: pediatric acute lymphoblastic leukemia; ALL; digital gene expression; second-generation sequencing; antisense transcript; alternative polyadenylation

INTRODUCTION

Acute lymphoblastic leukemia (ALL) is a heterogeneous disease that originates from lymphocyte progenitor cells of B- or T-cell origin. ALL comprises multiple distinct subtypes that are characterized by recurrent copy number alterations and structural chromosomal rearrangements, which have important clinical implications. Such cytogenetically distinct subtypes include B-cell precursor (BCP) leukemias with the chromosomal translocations t(12;21)[p13;q22][ETV6/RUNX1], t(9;22)[q11;q34][BCR/ABL1], dic(9;20)[p13;q11] and high hyperdiploidy (HeH) (>50 chromosomes) karyotypes. It is well established that ALL subtypes differ from a clinical perspective, but the underlying molecular consequences of most of the recurrent chromosomal abnormalities are poorly understood.

Expression microarrays have been applied for genome-wide expression analysis for classification of leukemia subtypes 1–3 and for identification of differentially expressed genes associated with drug resistance and treatment outcome in childhood ALL. 4 Although these studies provided new information regarding gene expression patterns in ALL, genes and affected pathways may have been missed because of technical limitations of hybridization-based methods. Sequencing-based methods like serial analysis of gene expression 5 generate absolute rather than relative measurements of gene expression without the bias of predesigned hybridization probes. Until recently, the cost and throughput of capillary sequencing technology have hindered their widespread use. The digital gene expression (DGE) profiling method, based on similar principles as serial analysis of gene expression, was made possible by the recent advances made in ‘second-generation’ DNA sequencing technologies 6–8. Today, millions of expression tags can be measured simultaneously at a fraction of the cost of capillary sequencing used in the original serial analysis of gene expression method. A major advantage of DGE, which overcomes the drawbacks of using predesigned hybridization probes as in microarray-based methods, is that it allows detection of expressed genes in a strand-specific manner. 9–11 Thus, the DGE method provides information on the polarity of the expressed transcripts, allowing detection of sense and antisense transcripts from the same gene region.

Antisense transcripts are endogenous RNA molecules that are transcribed from the noncoding DNA strand, on the opposite strand of a gene. Like protein-coding transcripts, antisense RNAs may contain a 5’ Cap and can be polyadenylated at the 3’ end. 12,13 Sense–antisense transcripts arising from the same gene region appear to be ubiquitously expressed in healthy and diseased mammalian cells. 11,13,14 For example, systematic investigation of antisense transcription by the FANTOM (Functional Annotation of the Mammalian Genome) consortium revealed that up to 72%
of all transcriptional units overlap with transcripts expressed from the non-coding DNA strand in human and mouse tissues. Sense-antisense transcript pairs have been observed in human lymphocytes, and using DGE of oligo-dT-captured RNA, the ratio between sense-antisense transcripts has been found to differ between cancer cells and normal matched tissues and between cancer subtypes. Focused studies on individual genes with hematopoietic and pro-leukemic functions have uncovered antisense RNAs that specifically modulate dosage of the sense transcript. Such antisense RNAs are likely to be involved in the regulation of many protein-coding genes, which may be important for leukemic progression. No genome-wide analysis of antisense transcripts in primary ALL cells has so far been conducted.

The presence of multiple DGE tags in a gene can be used to identify transcripts with 3' untranslated regions (UTRs) of different lengths. Widespread alternative polyadenylation (APA) in cancer cells and the preferential use of shorter 3' lengths. Widespread alternative polyadenylation (APA) in cancer transcript. Such antisense RNAs are likely to be involved in the regulation of many protein-coding genes, which may be important for leukemic progression. No genome-wide analysis of antisense transcripts in primary ALL cells has so far been conducted.

**MATERIALS AND METHODS**

**Patient samples**

Bone marrow or peripheral blood samples from 21 children with ALL representing five cytogenetic subtypes of ALL were included in this study (Table 1). Leukemic cells were isolated by 1.077 g/ml Ficoll-Isopaque (Pharmacia, Uppsala, Sweden) density-gradient centrifugation. The proportion of lymphoblasts was >90% as estimated by light microscopy of May-Grünwald-Giemsa-stained cytocentrifuge preparations. RNA was extracted from frozen cell pellets as described previously. RNA samples had an average RNA integrity number of 9.1 (minimum 7.8) according to Bioanalyzer analysis (Agilent Technologies, Santa Clara, CA, USA). The RNA was quantified by ultraviolet absorbance using a ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). The regional ethics committee approved this study, and patients and/or their guardians provided written informed consent. This study was conducted in accordance with the Helsinki Declaration.

**Preparation of sequencing libraries**

Sequencing libraries were prepared from 1 μg of total RNA using reagents from the Nall Digital Gene Expression Tag Profiling kit (Illumina Inc., San Diego, CA, USA). mRNA was captured on magnetic oligo(dT) beads and reverse transcribed into double-stranded cDNA (SuperScript II, Invitrogen, Carlsbad, CA, USA). The cDNA was cleaved using the restriction enzyme Nall. An adapter sequence containing the recognition sequence for the restriction enzyme MmeI was ligated to the Nall cleavage sites. The adapter-ligated cDNA was digested with MmeI to release the cDNA from the magnetic bead, while leaving 17 bp of sequence in the fragment. The fragments were dephosphorylated and purified by phenol-chloroform. A second adapter was ligated at the MmeI cleavage sites. Adapter-ligated cDNA fragments were amplified by PCR, and PCR products were purified on a 6% polacrylamide gel (Invitrogen). The ~96-bp PCR products were excised from the gel and eluted overnight, followed by ethanol precipitation and re-suspension (Illumina Inc.). Purified libraries were quality controlled and quantified on a Bioanalyzer using DNA 1000 series or High-Sensitivity chips (Agilent Technologies). DGE libraries were diluted to a 10 nM concentration and stored at −20 °C until sequencing.

**Sequencing and data processing**

Each DGE library was sequenced on an individual lane of a flow cell using an Illumina Genome Analyzer (GALII or GALIIx) for 18 cycles using reagents from version 2 cluster generation kits and version 3 sequencing kits (Illumina Inc.). Image analysis and base calling were performed using the Genome Analyzer pipeline v1.4. The first 17 bases of the tag sequences were extracted from the output

---

**Table 1. Clinical characteristics of ALL patients**

| Sample ID | Immunophenotype | Genetic subtype | Age at diagnosis (years) | WBC | Events |
|-----------|-----------------|-----------------|--------------------------|-----|--------|
| Patient 1 | Pre-B           | dic(9;20)       | 1.7                      | 93.7| CR1    |
| Patient 2 | Pre-B           | dic(9;20)       | 5.6                      | 77.0| CR1    |
| Patient 3 | Pre-B           | dic(9;20)       | 6.3                      | NA  | CR1    |
| Patient 4 | Pre-B           | HeH             | 3.3                      | 0.8 | CR1    |
| Patient 5 | Pre-B           | HeH             | 8.9                      | 95.0| CR1    |
| Patient 6 | Pre-B           | HeH             | 1.9                      | 23.3| Relapse|
| Patient 7 | Pre-B           | HeH             | 3.5                      | 11.2| CR1    |
| Patient 8 | Pre-B           | HeH             | 14.1                     | 5.0 | Relapse|
| Patient 9 | Pre-B           | HeH             | 3.1                      | 3.0 | CR1    |
| Patient 10 | Pre-B            | HeH             | 6                       | 24.5| CR1    |
| Patient 11 | Pre-B            | HeH             | 3.8                      | 39.6| Relapse|
| Patient 12 | Pre-B            | t(12;21)        | 3.1                      | 28.0| CR1    |
| Patient 13 | Pre-B            | t(12;21)        | 3.7                      | 12.3| CR1    |
| Patient 14 | Pre-B            | t(12;21)        | 4.7                      | 4.6 | SMN    |
| Patient 15 | Pre-B            | t(9;22)         | 10.7                     | 116.0| SMN    |
| Patient 16 | Pre-B            | t(9;22)         | 11.2                     | 64.4| CR1    |
| Patient 17 | Pre-B            | t(9;22)         | 13.5                     | 130.0| Relapse|
| Patient 18 | T-ALL            | T-ALL           | 13.9                     | 139.0| CR1    |
| Patient 19 | T-ALL            | T-ALL           | 13.4                     | 66.5| Resistant disease |
| Patient 20 | T-ALL            | T-ALL           | 10.8                     | 234.0| DCR1  |
| Patient 21 | T-ALL            | T-ALL           | 17.2                     | 126.0| CR1    |

Abbreviations: ALL, acute lymphoblastic leukemia; CR1, continuous first remission; DCR1, dead in CR1; NA, not available; SMN, secondary malignant neoplasm; T-ALL, T-cell lineage ALL; WBC, white blood cell. *Dic(9;20), dicentric chromosome (9;20)(p13;q11): HeH, high hyperdiploidy; t(12;21), translocation between chromosomes (12;21)(p13;q22)(ETV6/RUNX1); t(9;22), translocation between chromosomes (9;22)(q11;q34)(BCR/ABL1); T-ALL, T-cell ALL. ALL diagnosis was established by analysis of morphology, immunophenotype and cytogenetics of leukemic cells. Fluorescence in situ hybridization and/or reverse-transcriptase PCR were applied to identify t(12;21) and dic(9;20) rearrangements. **WBC count at diagnosis (10^9 cells/l).
files using a stringent base quality cutoff equivalent to a phred score of 20, discarding tags if they had any base with a score below 20. Unique tags were sorted and counted in each of the DGE libraries using custom Perl scripts written for DGE analysis.

Annotation of sequenced tags
DGE tags were annotated to the human transcriptome (Ensembl version 58) by mapping the reads to the sequence flanking NlaIII restriction sites on both coding and non-coding strands. Tags matching more than one gene region were discarded. Tag counts were normalized to tags per million (TPM) by dividing the raw tag count by the total number of tags from each library and multiplying by one million. The total expression profile for each gene was calculated by summing all tags mapped to the same gene, including intronic tags (Supplementary Information). DGE data are available online at the Gene Expression Omnibus under accession number GSE26878. Previously reported sense/antisense expressed sequence tags or mRNAs (n = 8652) were downloaded from the Natural Antisense Transcripts Database (NATSDB, release 2006/2007). A total of 8554 pairs remained after conversion to Genome Reference Consortium human build 37 (GRCh37). All uniquely annotated antisense tag sequences from DGE were examined to assess whether they are located within a region flanked by previously observed pairs in NATSDB.

Identification and analysis of APA
Genes with at least two uniquely annotated tags in the 3' UTR of the last exon or within 1000 bp downstream of the 3' gene boundary were marked as potentially affected by APA. To compare the APA detected by DGE with predicted polya cleavage sites, we used the Alternative Splicing and Transcript Diversity database (release 1.1 build 9), which contains 41 024 APA sites. A total of 41 005 sites remained after conversion to GRCh37. To investigate the presence of micro-RNA (miRNA) target sites in the genes potentially affected by APA, 54 199 predicted seed regions were downloaded from the TargetScan Human database (release 5.1). Omitting tags with abundance below 2 TPM, tags that mapped to more than one gene and tags that had no match with the reference transcripts in the Ensembl database, we obtained 20-46 thousand unique nucleotide sequence tags per library that mapped to the transcriptome (Figure 1a, Supplementary Table S1). Omitting tags with abundance below 2 TPM, tags that mapped to more than one gene and tags that had no match with the reference transcripts in the Ensembl database, we obtained 20-46 thousand unique nucleotide sequence tags per library that mapped to the transcriptome (Figure 1a, Supplementary Table S1). In total, we observed a robust expression of 17 313 genes with tags ≥2 TPM in the 21 libraries from ALL cells. The dynamic range for the gene expression measured by DGE was broad, ranging from 2-7.3 x 10^3 TPM (Figure 1b), which corresponds to 0.7-2.2 x 10^4 transcripts per cell, assuming that one transcript per cell is roughly equal to one transcript in a 3 x 10^5.
sets, we included two additional expression data sets. Taken
expression data sets for differentially expressed genes detected
antisense direction.

different bins of expression levels. The expression levels are in tags
per million (TPM) on a log2-transformed scale on the horizontal
(Figure 3).

PCR and expressed tag count determined by DGE were consistent
by NSC were validated by quantitative PCR. For each gene and
identified by NSC (Figures 2a and b). Five of the genes identified
by NSC were validated by quantitative PCR. For each gene and
subsubtype, the relative expression levels determined by quantitative
PCR and expressed tag count determined by DGE were consistent
(Figure 3).

We used the Oncomine tool33 to mine published gene expression data sets for differentially expressed genes detected here by NSC of the DGE data. In addition to the Oncomine data sets, we included two additional expression data sets. Taken
together, expression data from >1600 ALL patients validated the
differential expression of 17 of the 20 genes between the BCP ALL
and T-ALL subtypes (Supplementary Table S4). In all, 18 of 34 BCP
ALL subtype-specific genes were validated in ~750 ALL samples
(Supplementary Table S5). The overlap of differentially expressed genes between our relatively small number of samples and genes reported by multiple studies of large ALL sample sets and provides evidence for the robustness and biological relevance of the genes identified here by DGE in combination with NSC.

The 54 genes selected by the 2 classifiers are enriched for
higher-order molecular and cellular functions such as cell
signaling (n = 7, P < 0.001), cellular assembly and organization
(n = 7, P < 0.05), as well as cell death (n = 14, P < 0.0026), which are
relevant for the ALL phenotype. Four of the genes (namely ASS1,
GAB1, SOX11 and THBS1) are known as diagnostic or prognostic
markers for other hematological malignancies,34-37 but have not
previously been studied as markers for ALL. Pairwise analyses of
the DGE data from the four subtypes of BCP ALL distinguished
differentially expressed genes, which point at distinct pathways in
HeH, t(9;22), t(12;21) and dic(9;20) ALL (Supplementary Table S6).
The p53 signaling pathway was highlighted in cells of the dic(9;20)
subtype, genes from the interferon signaling and innate immunity
systems were highlighted in t(9;22) cells, differential regulation of
anti-apoptosis genes was highlighted in t(12;21) cells and
apoptosis signaling in HeH cells (Table 2).

Quantification of antisense transcripts in ALL cells
We took advantage of the strand specificity of DGE to measure
antisense expression in ALL cells. We detected a total of 30150
antisense tags at a level ≥2 TPM, which corresponds to antisense
transcription of 48.6% of the expressed gene regions (Figure 1a).
The correlation between the antisense expression levels measured
across all gene regions in replicate libraries from the same RNA
sample was high (r = 0.70, Figure 4a). More than 30% of the
antisense tags observed in ALL cells were located within the
borders of a previously described sense/antisense transcript pair
in the Natural Antisense Transcript database.24 The antisense
transcripts were expressed at substantial levels (2–1.7 × 10^3 TPM),
but in most cases, they were less abundant than sense transcripts
(Figure 1b). However, we observed higher antisense expression
levels for 6.5% (549/8405) of the sense/antisense gene pairs. The
correlation between expression levels of sense and antisense tags
originating from the same gene region varied over a broad range,
from perfectly correlated (r = 1.0) to strongly anti-correlated
(r = –0.7) pairs (Figure 4b). The absence of a systematic
correlation between reciprocal sense-antisense pairs indicates
that transcription from the antisense strand is either concordant
with transcription in the sense orientation or discordant. This is
consistent with observations in other studies of different
tissues.38,39 In addition, 173 genes were detected only by
antisense tags.

Classification of ALL subtypes by antisense transcripts
We performed supervised classification with NSC to identify gene
regions with subtype-specific antisense transcripts. NSC defined
19 antisense tags expressed in 15 gene regions that distinguished
the T-ALL and BCP ALL subtypes (P < 0.001) (Supplementary Table S7)
and 83 antisense tags expressed in the antisense orientation
of 76 gene regions (P < 0.001) that distinguished the 4 BCP ALL
subtypes (Supplementary Table S8). Analogously to genes
expressed in the sense direction (Figures 2a and b), hierarchical
clustering of samples by antisense tags selected by the classifier
resulted in perfect separation of the samples according to their
respective subtypes (Figures 2c and d). The majority of gene
regions highlighted by NSC classification of antisense tags were
different from those highlighted by expression in the sense
direction. An exception is the SOX11 gene that was identified as

![Figure 1. Distribution between gene locations and abundance of annotated tags. To annotate the sequenced tags, we created a database of all possible 17-bp tag sequences next to an NlaIII site from the Ensembl transcriptome database. (a) The bars show the proportion of the total number of tags by location in gene regions. The canonical location refers to the tag originating from the most 3' NlaIII site in a gene, to which 41% mapped. Overall, 22% of the tags were mapped to exons, which may represent transcript isoforms that are not listed in Ensembl. In all, 3.4% of the tags, which presumably originate from unprocessed pre-mRNA transcripts, mis-spliced transcripts, or unannotated exons, were annotated to intronic gene sequences. Overall, 3.6% of the sequenced tags mapped in an antisense orientation to gene regions in the Ensembl database. The remaining tags had multiple annotations (23%) or were not found in our database of possible tag sequences (7%). No significant difference in the distribution between gene locations was observed for the annotated tags between the ALL subtypes (data not shown). (b) The bars show the proportion of annotated tags at different bins of expression levels. The expression levels are in tags per million (TPM) on a log2-transformed scale on the horizontal axis. Black bars indicate tags annotated to genes in the sense direction and light gray bars indicate tags annotated to genes in the antisense direction.](https://example.com/figure1.png)
Figure 2. Hierarchical clustering of ALL samples according to gene regions defined by nearest shrunken centroid (NSC) classification based on DGE data from sense and antisense transcripts. (a) Heatmap of the expression levels of the 20 genes defined by the NSC classifier for discriminating between the samples of BCP and T-ALL subtypes. These genes are listed in Supplementary Table S2. (b) Heatmap of the expression levels of 34 genes defined by the NSC classifier for discriminating between the 4 BCP subtypes. The 34 genes are listed in Supplementary Table S3. (c) Heatmap of the expression levels of 19 antisense expressed tags defined by the NSC classifier for discriminating the samples with BCP and T-ALL subtypes. The tags and corresponding gene regions are listed in Supplementary Table S7. (d) Heatmap of the expression levels of 83 antisense expressed tags defined by the NSC classifier for discriminating the BCP subtypes, the tags and corresponding gene regions are listed in Supplementary Table S8. ALL samples are shown in horizontal rows, and genes are shown in vertical columns. The color code for the BCP vs T-ALL comparisons (panels a and c) is shown in the upper right corner of panel a. BCP samples are indicated in light gray and T-ALL samples are indicated in dark gray. The color code for the BCP subtypes comparisons (panels b and c) is shown in the upper right corner of panel b as follows: t(9;22) samples (n = 3) are indicated in light pink, HeH samples (n = 8) are indicated in light blue, t(12;21) samples (n = 3) are indicated in dark yellow and dic(9;20) samples (n = 3) are indicated in dark blue. The color code for the gene expression levels as number of transcripts per million (TPM) is shown on the right-hand side of each Heatmap.
expressed specifically in the t(12;21) subtype samples in both sense and antisense orientations. SOX11, a non-B-cell lineage transcription factor encoded by an intron-less gene (Figure 5), is involved in the regulation during embryonic development and in the central nervous system. The SOX11 protein is expressed in a subset of mantle cell lymphoma patients, in whom elevated expression marks better overall survival. Genes highlighted by antisense tags and subtype-specific sense genes belong to similar functional categories, like hematological system development ($n = 20, P < 0.049$) and cell death ($n = 18, P < 0.037$). Notably, antisense transcription of the important hematological regulators NOTCH3 and PAX5 was observed in T-ALL cells and BCP ALL cells, respectively. Sense and antisense transcription from the SOX11 and PAX5 gene regions were subsequently validated by strand-specific reverse transcription and PCR (Supplementary Figure S3, Supplementary Table S9, Supplementary Materials and methods). The result for antisense transcription at the NOTCH3 locus was inconclusive.

In light of recent findings of sense/antisense ratios that differ between normal and malignant tissues, we investigated...
DISCUSSION

In this study, we determined the genome-wide gene expression profiles of 21 primary ALL cell samples by ‘second-generation’ sequencing of short cDNA tags. We demonstrate that this unbiased, sequencing-based approach not only allows precise genome-wide expression profiling but also provides novel information on gene expression in ALL cells by detecting antisense transcripts and APA.

The pathway-enrichment results from the dic(9;20) subtype are particularly interesting. Although dic(9;20) is an established recurrent chromosomal aberration with an estimated prevalence of ~5% of ALL cases in the Nordic countries,43 no genome-wide expression studies have been performed on this subtype. No fusion

**Table 2. Pathway enrichment of genes with differential expression between BCP ALL subtypes**

| Subtype | IPA | P-value | Pathway-associated genes\* \( \text{(upregulated/downregulated)} \) |
|---------|-----|---------|-------------------------------------------------------------|
| dic(9;20) | TP53 signaling | 0.032 | CDKN2A(−), CSNK1D(+), CTNNB1(+) |
|         | Interferon signaling | 0.000089 | IFIT1(+), IFIT3(+), IRF9(+), OAS1(+), STAT1(+), STAT2(+) |
|         | Activation of IRF by cytoplasmic pattern recognition receptors | 0.00034 | DHX58(+), IFIT2(+), IRF9(+), NFKB2(+), NFKBIE(+), STAT1(+), STAT2(+) |
|         | Role of RIG1-like receptors in antiviral innate immunity | 0.002 | CASP8(+), CASP10(+), DHX58(+), NFKB2(+), NFKBIE(+) |
| t(9;22) | Anti-apoptosis | 0.0026 | BAG3(+), BCL2L1(+), BIRC7(+), CD27(+) |
| HeH | VDR/RXR activation | 0.014 | KLF4(+), NCOR2(+), PDGFA(−), PRKCQ(+) |
|        | Apoptosis signaling | 0.00081 | BCL2L1(−), CAPN2(+), CAPN3(−), CASP8(−), CASP10(−), NFKB2(−), NFKBIE(−), NFKBIV(−), PRKCQ(−) |
|        | OX40 signaling | 0.004 | BCL2L1(−), HLA-DOB(−), NFKB2(−), NFKBIE(−), TNFSF4(−) |

Abbreviations: ALL, acute lymphoblastic leukemia; IPA, ingenuity pathway analysis; IRF, interferon-regulatory factor; RIG1, retinoic acid inducible gene-1; VDR/RXR, vitamin D3 receptor/retinoic X receptor. *IPA canonical pathways or IPA tox lists. †Fisher’s exact test P-value. ‡Gene symbol according to HUGO Gene Nomenclature Committee (http://www.genenames.org/). (+) Indicates significant upregulation. An increasing number of (+) signs indicates stronger upregulation. (−) Indicates significant downregulation. An increase in the number of (−) signs indicates stronger downregulation.

**Figure 4.** Detection of antisense transcription by digital gene expression (DGE). (a) Positive correlation between expression levels of antisense tags in DGE replication experiments from the same ALL patient RNA sample (Pearson’s correlation coefficient \( r = 0.70 \)). (b) Lack of systematic correlation between sense and antisense expression from the same gene loci. Each dot represents one gene. Pearson’s correlation coefficients for each gene with sense and antisense transcription are plotted on the y axis.

whether the ratio differed between the four subtypes of BCP ALL or between all the BCP ALL cells as one group and T-ALL cells. Although we observed a subset of 272 genes with differences in the sense/antisense ratios between subtypes, this result did not reach statistical significance \( P < 0.05 \) after multiple testing correction (data not shown).

Use of APA in ALL cells

Alternative polyadenylation (APA) changes the length of the 3’ UTR of genes, which leads to the expression of mRNA isoforms of different lengths. If the isoform length differs by at least one Nlal site, these different transcripts are detectable in the DGE data as multiple expressed 3’ tags in a gene region. We detected signs of APA in 38.2% (6619) of the expressed genes. Of the pairs of consecutive expressed tags in the 3’ UTR of the last exon or flanking region, 17% were supported by the presence of a predicted polyA cleavage site between the tags.20 One of the functional consequences of shorter 3’ UTRs is increased stability of a transcript as a result of the loss of miRNA-binding sites.20 Using available data on predicted miRNA target sites,26 we found that genes with APA contained significantly more miRNA target sites than did genes without evidence of APA \( (P < 1 \times 10^{-15}) \), suggesting that APA in ALL cells may have a major effect on miRNA-mediated regulation of gene expression. Interestingly, we observed that genes with APA are also more likely to have antisense tags in the same gene region \( (P < 1 \times 10^{-17}) \). These findings suggest that APA, the presence of miRNA-binding sites and antisense transcripts may be spatially or functionally connected and important for gene regulation in ALL cells.

We did not observe subtype-specific patterns for genes with APA. Instead, when analyzing the 1687 genes with exactly two tags in the 3’ UTR or flanking region, we found 148 genes with preferred short 3’ UTRs and 679 genes with preferred longer 3’ UTRs. The genes with preferred shorter 3’ UTRs compared with the genes with preferentially longer 3’ UTRs are enriched for functions in cell-cycle control, cellular assembly and organization, DNA replication, recombination and repair \( (P < 0.05) \). For a gene list and functional annotation of the genes with shorter 3’ UTRs, see Supplementary Table S10.
Digital gene expression in leukemic cells
J Nordlund et al

The results from our study provide a basis for further functional studies of regulatory RNA molecules that may affect the malignant transformation of precursor B cells into ALL cells. Antisense transcripts are remarkably abundant in primary ALL cells, with antisense transcription of ~49% of the expressed genes. More than 30% of the antisense transcripts identified here have been observed previously, indicating that the antisense transcriptome in ALL overlaps with that in other cell types. Many of the antisense loci identified here are reported for the first time, presumably because antisense transcription has not been previously investigated on a genome-wide scale in primary ALL cells. As we used oligo-dT capturing and many RNAs lack polyA tails, we most likely underestimate the actual contribution of antisense transcripts in ALL cells. Aberrantly expressed antisense transcripts may contribute to disease by inducing chromosomal changes, through DNA–RNA interactions, or through transcription interference. For example, the tumor-suppressor gene PT5 has an antisense RNA that silences the P15 gene through epigenetic alteration of heterochromatin in leukemia cell lines. We observed antisense expression from genes known to have critical roles in cell death and regulation of gene expression in leukemia, such as the PAX5 and NOTCH3 genes. PAX5 is a target of somatic mutations in BCP ALL, and the Notch pathway is frequently activated in T-ALL.

Both these genes are lineage-specific regulators at early stages of B/T-cell development. In addition, SOX11 has been implicated as an important growth regulator in hematological malignancies. In our study, we also observed extensive APA and a strong association between genes with APA and antisense expression in the same gene region (P < 1 × 10^-15). The extent and regulatory function of antisense transcription and APA in ALL may be clarified in greater detail by strand-specific transcriptome sequencing, which has become feasible only recently.

Complete transcriptome sequencing will be highly informative in future studies of ALL and other malignancies because it provides information on transcript sizes, isoforms and expression of fusion genes, but DGE has some important practical advantages over transcriptome sequencing. First, DGE requires less RNA than most other transcriptome sequencing methods, and second, the computational analysis of DGE data is less challenging. Taking advantage of the increasing capacity of 'second-generation' sequencing technologies, DGE could easily be scaled up using indexing to allow inexpensive and rapid digital expression profiling in large sample collections.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ACKNOWLEDGEMENTS
NGS was performed at the SNP&SEQ Technology Platform in Uppsala (http://www.sequencing.se) with the assistance of Ulrika Liljedahl, Marie Lindensson and Kristina Larsson. Affymetrix data were generated at the Uppsala Array Platform (http://www.medsci.uu.se/kinfarm/arrayplatform). Both platforms are supported by the Uppsala University, Uppsala University Hospital and the Science for Life Laboratory, Uppsala. The SNP&SEQ Platform is supported by the Swedish Research Council for Infrastructures (numbers 70374401 and 80576801). We thank our colleagues in the Nordic Society of Pediatric Hematology and Oncology and patients who provided the samples. We also thank Lili Milano for assistance with the experimental design, Annabett H Petersen for helpful discussions, Anders Lundmark for statistical expertise, Mårten Fryknäs for assistance with Oncomine and Matilda Canderby for technical assistance. This work was funded by grants from the Swedish Cancer Society, the Swedish Childhood Cancer Foundation, the Swedish Research Council for Science and Technology (90559401) and the Swedish Foundation for Strategic Research (RBC8-008).

Data are available at the Gene Expression Omnibus under SuperSeries GSE26878 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26878).

REFERENCES
1 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP et al. Molecular classification of cancer: discovery and class prediction by gene expression monitoring. Science 1999; 286: 531 - 537.
2 Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahoulou R et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 2002; 1: 133 - 143.

3 Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 2003; 102: 2951 - 2959.

4 Hollman A, Cheuk MH, den Boer ML, Yang W, Veeman AJ, Kazemier KM et al. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. N Engl J Med 2004; 351: 533 - 542.

5 Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995; 270: 484 - 487.

6 Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008; 24: 133 - 141.

7 Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet 2010; 11: 31 - 46.

8 Nielsen KL, Hogh AL, Emmerssen J. DeepSAGE-digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic Acids Res 2006; 34: e133.

9 Wu ZJ, Meyer CA, Choudhury S, Shipstain M, Maruyama R, Bessarabaova M et al. Gene expression profiling of human breast tissue samples using SAGE-Seq. Genome Res 2010; 20: 1730 - 1739.

10 Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX et al. The Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Leukemia (2012) 1218 - - 1227.

11 T flatou et al. 2005; 484 - 487.

12 Mardis ER. The impact of next-generation sequencing technology on genetics. Science 2006; 312: 68 - 74.

13 The FANTOM Consortium Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC et al. The transcriptional landscape of the mammalian genome. Science 2005; 309: 1555 - 1562.

14 Zhang Y, Li J, Kong L, Gao G, Liu QR, Wei L. NATsDB: Natural Antisense Transcripts Dataset. Nucleic Acids Res 2007; 35: D156 - D161.

15 Zachariadis V, Gauffin F, Kuchinskaya E, Heyman M, Schoumans J, Blennow E et al. Pelosi (2010; 133 - 141).

16 Schoumans J, Shipstain M, Choudhury S, Wu Z, Protopopov A, Yao J et al. Altered antisense-to-sense transcript ratios in breast cancer. Proc Natl Acad Sci USA 2010; e-pub ahead of print 22 November 2010.

17 Zachariaidou V, Gaffin F, Kuchinskaya E, Heyman M, Schoumans J, Bienen E et al. The frequency and prognostic impact of dic(9;20)(p13.2;q11.2) in childhood precursor acute lymphoblastic leukemia: results from the NOPHO ALL-2000 trial. Leukemia 2011; 25: 622 - 628.

18 Schoumans J, Johannson B, Corcoran M, Kuchinskaya E, Golovleva I, Grander D et al. Characterisation of dic(9;20)(p11.13;q11) in childhood B-cell precursor acute lymphoblastic leukaemia by tiling resolution array-based comparative genomic hybridisation reveals clustered breakpoints at 9p13.2 and 20q11.2. Br J Haematol 2006; 134: 492 - 499.

19 Sulong S, Moorman AV, Irving JA, Strefford JC, Konn ZJ, Case MC et al. Clinical and cytogenetic features of pediatric dic(9;20)(p13.2;q11.2)-positive B-cell precursor acute lymphoblastic leukaemias: a Nordic series of 24 cases and review of the literature. Genes Chromosomes Cancer 2008; 47: 149 - 158.

20 Zhang Y, Xiong Y, Yarbrough WG. ARF promotes MM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the Rb and p53 tumor suppression pathways. Cell 2002; 109: 723 - 734.
49 Bungaro S, Dell’Orto MC, Zangrando A, Basso D, Gorletta T, Lo Nigro L et al. Integration of genomic and gene expression data of childhood ALL without known aberrations identifies subgroups with specific genetic hallmarks. Genes Chromosomes Cancer 2009; 48: 22 - 38.

50 Shaulian E, Schreiber M, Piu F, Beeche M, Wagner EF, Karin M. The mammalian UV response: c-Jun induction is required for exit from p53-imposed growth arrest. Cell 2000; 103: 897 - 907.

51 Lecomte S, Desmots F, Le Masson F, Le Goff P, Michel D, Christians ES et al. Roles of heat shock factor 1 and 2 in response to proteasome inhibition: consequence on p53 stability. Oncogene 2010; 29: 4216 - 4224.

52 Weaver BAA, Cleveland DW. Does aneuploidy cause cancer? [Curr Opin Cell Biol 2006, 18: 658-667]. Curr Opin Cell Biol 2007; 19: 246 - 246.

53 Passalacqua M, Patrone M, Sparatore B, Melloni E, Pontremoli S. Protein kinase C- theta is specifically localized on centrosomes and kinetochores in mitotic cells. Biochem J 1999; 337 (Part 1): 113 - 118.

54 Chattopadhyay A, Chiang CW, Yang E. BAD/BCL-X(L) heterodimerization leads to bypass of G0/G1 arrest. Oncogene 2001; 20: 4507 - 4518.

55 Boise LH, Gonzalez-Garcia M, Postema CE, Ding L, Lindsten T, Turka LA et al. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 1993; 74: 597 - 608.

56 Zhang Y, Liu XS, Liu QR, Wei L. Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. Nucleic Acids Res 2006; 34: 3465 - 3475.

57 Faghihi MA, Wahlestedt C. Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol 2009; 10: 637 - 643.

58 Coyaud E, Struski S, Prade N, Familiades J, Eichner R, Quelen C et al. Wide diversity of PAX5 alterations in B-ALL: a Groupe Francophone de Cytogenetique Hematologique study. Blood 2010; 115: 3089 - 3097.

59 Aifantis I, Raetz E, Buonamici S. Molecular pathogenesis of T-cell leukaemia and lymphoma. Nat Rev Immunol 2008; 8: 380 - 390.

Supplementary Information accompanies the paper on the Leukemia website (http://www.nature.com/leu)