

ORIGINAL ARTICLE

Digital gene expression profiling of primary acute lymphoblastic leukemia cells

J Nordlund¹, A Kiialainen^{1,9}, O Karlberg¹, EC Berglund¹, H Göransson-Kultima², M Sønderkær³, KL Nielsen³, MG Gustafsson², M Behrendtz⁴, E Forestier⁵, M Perkkiö⁶, S Söderhäll⁷, G Lönnerholm⁸ and A-C Syvänen¹

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 signaling 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 ($P < 1 \times 10^{-15}$). 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 expression⁵ 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%

¹Molecular Medicine, Department of Medical Sciences, Uppsala University, Uppsala, Sweden; ²Department of Medical Sciences, Cancer Pharmacology and Computational Medicine, Uppsala University, Uppsala, Sweden; ³Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, Aalborg, Denmark; ⁴Department of Pediatrics, Linköping University Hospital, Linköping, Sweden; ⁵Department of Medical Biosciences, University of Umeå, Umeå, Sweden; ⁶Department of Pediatrics, Kuopio University Hospital, Kuopio, Finland; ⁷Childhood Cancer Research Unit, Department of Women and Child Health, Astrid Lindgren Children's Hospital, Karolinska University Hospital, Stockholm, Sweden and ⁸Department of Women's and Children's Health, Uppsala University, Uppsala University Children's Hospital, Uppsala, Sweden. Correspondence: Dr A-C Syvänen, Department of Medical Sciences, Molecular Medicine, Academic Hospital, Uppsala University, Entrance 70, Third Floor, 751 85 Uppsala, Sweden. E-mail: Ann-Christine.Syvanen@medsci.uu.se

⁹Current address: Pharma Research and Early Development, F. Hoffmann-La Roche Ltd, Basel, Switzerland.

Received 25 February 2011; revised 14 October 2011; accepted 4 November 2011; published online 16 December 2011

of all transcriptional units overlap with transcripts expressed from the non-coding DNA strand in human and mouse tissues.^{15,16} Sense-antisense transcript pairs have been observed in human lymphocytes,¹⁴ 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.¹¹ Focused studies on individual genes with hematopoietic and pro-leukemic functions have uncovered antisense RNAs that specifically modulate dosage of the sense transcript.^{17,18} 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' UTRs in proliferating cells have been observed.^{19,20} Furthermore, antisense transcripts and APA seem to be correlated.^{21,22} The DGE method can detect antisense transcripts and APA and could thus be an excellent tool for investigation of these events in ALL cells. In this study, we applied the DGE method for classification analysis of ALL subtypes and to detect subtype-specific antisense expression and APA in ALL cells.

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 $\geq 90\%$ as estimated by light microscopy of May-Grünwald-Giemsa-stained cytocentrifugate preparations. RNA was extracted from frozen cell pellets as described previously.²³ 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 μ g of total RNA using reagents from the *Nlalll* 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 *Nlalll*. An adapter sequence containing the recognition sequence for the restriction enzyme *MmeI* was ligated to the *Nlalll* 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% polyacrylamide 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 (GAII or GAIIx) 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 ^a	Age at diagnosis (years)	WBC ^b	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. ^aDic(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. ^bWBC 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).²⁴ 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.²⁵ 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),²⁶ 52 913 of which remained after conversion to GRCh37.

Quantitative PCR

The expression levels of five genes determined by DGE were validated by quantitative PCR with TaqMan gene expression assays with FAM dye-labeled minor groove binding probes (Applied Biosystems, Carlsbad, CA, USA). cDNA was synthesized from 200 ng of total RNA using the High-Capacity RNA-to-cDNA kit (Applied Biosystems). Five ng of cDNA was used in each reaction performed on a StepOnePlus Real-Time quantitative PCR system (Applied Biosystems). Each cDNA was analyzed in triplicate and the average Ct-value (ΔCt) was calculated per sample. Relative expression levels were calculated with the $2^{\Delta\Delta Ct}$ method.²⁷ *YWHAH* was used as an endogenous control, as it was the housekeeping gene with the lowest s.d. in the DGE data.

Statistical analyses

Statistical analyses were performed in R using tools from Bioconductor.²⁸ Hierarchical clustering was performed by conventional agglomerative clustering ('hclust') with one minus the correlation coefficient as the distance measure for pairs of individuals and average linkage as the measure of cluster similarity. The gene lists were analyzed by ingenuity pathway analysis (Ingenuity Systems, Redwood City, CA, USA). Ingenuity pathway analysis *P*-values were calculated using Fisher's exact test by measuring the difference in proportion overlapping genes for each function or pathway between the genes highlighted by a particular analysis and the total number of genes measured by DGE (17 199 genes were pathway eligible). Pairwise comparisons were performed with the software tool for empirical analysis of

DGE in R (edgeR).²⁹ Genes were called as differentially expressed when their adjusted *P*-values were < 0.05 . *P*-values were adjusted for multiple testing by the Benjamini and Hochberg approach. For multivariate analysis of gene expression, the method of nearest shrunken centroid (NSC) classifiers³⁰ as implemented in the R package 'pamr' was used. *P*-values for the performance of the classifiers were calculated by comparing the number of CV errors in the real data with the number of CV errors in the shuffled data, in which the subtype labels of the samples were randomly permuted 1000 times.³¹ The NSC scores reported here for each subtype are also known as the shrunken differences,³⁰ which is a standardized difference between the values of a given gene in the global centroid and the shrunken (modified) centroid for each subtype. For more details, see Supplementary Information.

RESULTS

Annotation of sequence tags

Taking advantage of the precise (digital) determination of gene expression levels by 'second-generation' tag sequencing (DGE), we generated expression profiles of cells taken at diagnosis from 17 patients with BCP ALL of four subtypes: high hyperdiploidy (HeH) ($n = 8$), t(9;22) *BCR-ABL1* ($n = 3$), t(12;21) *ETV6-RUNX1* ($n = 3$), dic(9;20) ($n = 3$) and four patients with T-cell lineage ALL (T-ALL) (Table 1). We prepared DGE sequencing libraries using *NlaIII* as the anchoring enzyme, and sequenced each of these libraries using an Illumina Genome Analyzer, which yielded 4.9–23.7 million quality-filtered sequence reads (tags) per sample (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 $2-7.3 \times 10^4$ TPM (Figure 1b), which corresponds to $0.7-2.2 \times 10^4$ transcripts per cell, assuming that one transcript per cell is roughly equal to one transcript in 3×10^5 .³² Replicate assays from different stages of the DGE procedure confirmed the reproducibility of the expression measurements using DGE (Supplementary Figure S1).

To benchmark DGE against the most commonly used method for gene expression analysis, we compared the DGE levels of 12 RNA samples with expression levels from Affymetrix GeneChips (Affymetrix, Santa Clara, CA, USA). Correlation between the gene expression levels measured by the two platforms was $\rho = 0.61$, indicating that DGE profiles are comparable to those from the Affymetrix GeneChips. DGE detected 22% more genes ($n = 3251$) with low average expression levels of ~ 10.7 TPM and had a 15-fold wider dynamic range than hybridization on GeneChips (Supplementary Information and Supplementary Figure S2).

Utility of DGE and NSC to classify ALL subtypes

We applied supervised classification by NSC³⁰ to obtain DGE signatures that are characteristic of ALL subtypes. NSC classification selected a set of 20 genes to distinguish between the BCP and T-ALL samples (Supplementary Table S2) and a set of 34 genes to distinguish between the BCP subtypes HeH, t(9;22), t(12;21) and dic(9;20) (Supplementary Table S3). Applying NSC for comparing BCP ALL cells against T-ALL cells resulted in an average of 1/21 cross-validation errors ($P < 0.001$), and the multivariate analysis of BCP ALL subtypes resulted in an average of 2.6/17 cross-validation errors ($P < 0.001$), indicating that the classifier did not classify the samples by chance. Evidence for the robustness of DGE was provided by perfect separation between the ALL subtypes by hierarchical clustering of samples according to gene sets



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.

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

We used the OncoPrint tool³³ to mine published gene expression data sets for differentially expressed genes detected here by NSC of the DGE data. In addition to the OncoPrint 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 30 150 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 ($2-1.7 \times 10^3$ 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

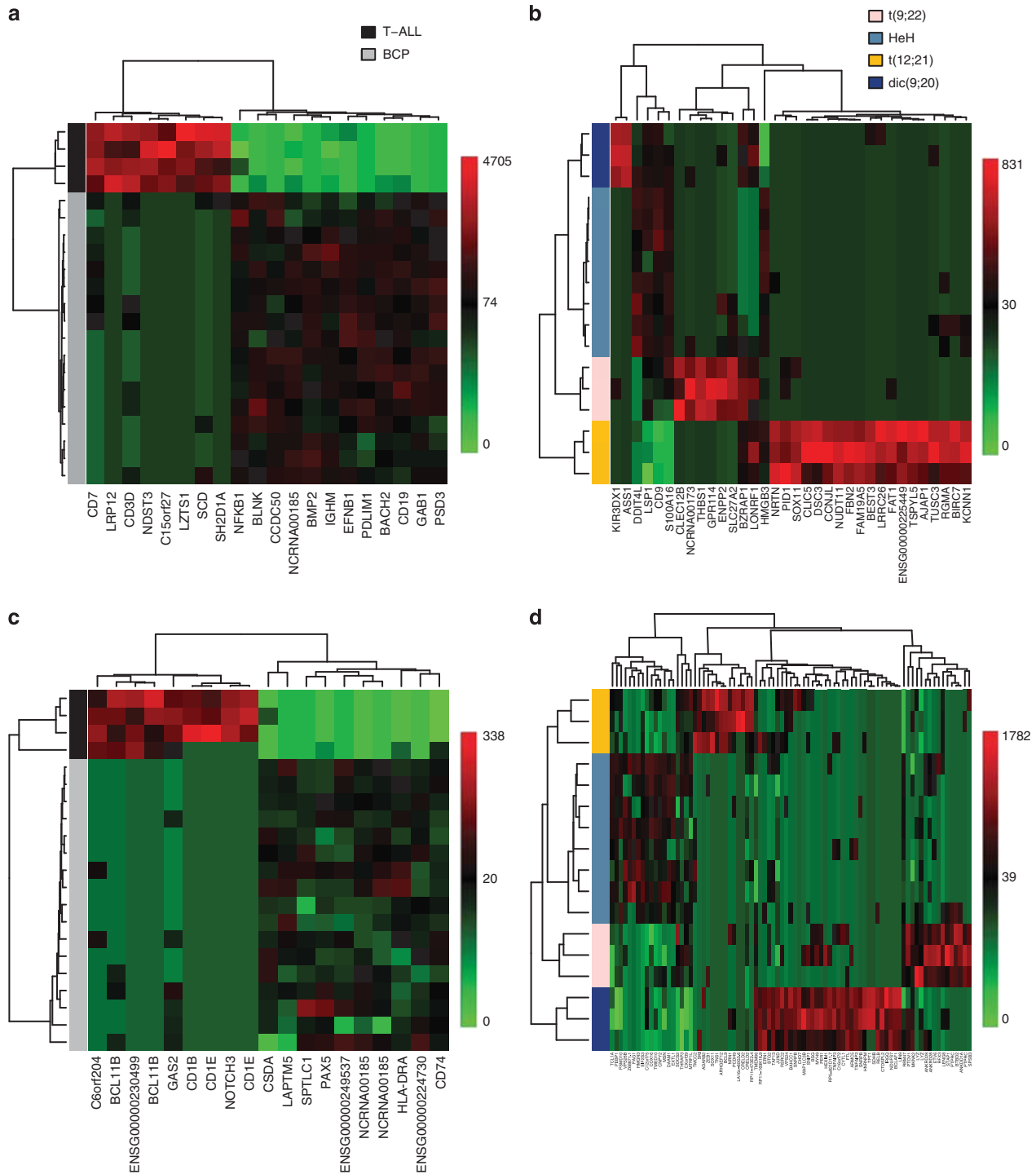


Figure 2. Hierarchical clustering of ALL samples according to gene regions defined by nearest shrunken centroid (NSC) classification based on DGE data from sense and antisense transcripts. **(a)** Heatmap of the expression levels of the 20 genes defined by the NSC classifier for discriminating between 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.

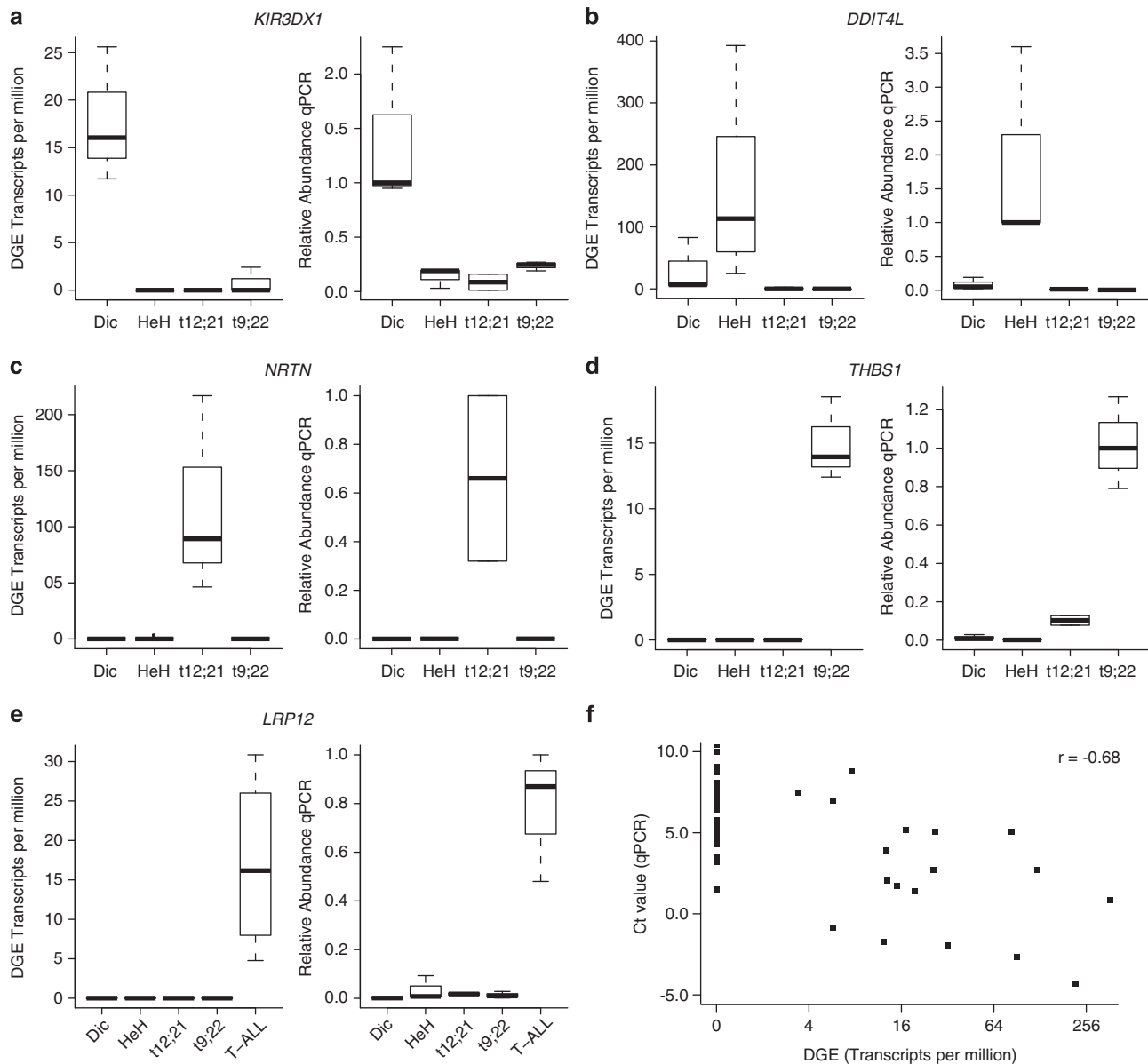


Figure 3. Validation of digital gene expression (DGE) levels by quantitative PCR (qPCR). The RNA expression levels of five genes (a) *KIR3DX1*, (b) *DDIT4L*, (c) *NRTN*, (d) *THBS1* and (e) *LRP12* highlighted in the NSC classification analysis were determined by qPCR. Three dic(9;20), t(9;22), HeH and T-ALL and two t(12;21) samples were analyzed by qPCR (see Table 1 for patients). The relative amounts of qPCR products are plotted in each panel to the right and the DGE measurements in transcripts per million (TPM) are plotted to the left. For each qPCR, one of the samples from the relevant subtype was selected as the reference for normalization. The relative expression levels measured by qPCR for the remaining samples were calculated in relation to this sample. The error bars show the maximum and minimum relative expression levels observed for each subtype. The DGE value in transcripts per million (TPM) for each gene in each sample is plotted by subtype on the left side of each panel. The error bars show s.d. for the TPM values. (f) Correlation between the expression levels measured for each of the five genes in each RNA sample using DGE and qPCR. The Δ Ct values (quantity of target gene/quantity of endogenous control) from qPCR are plotted on the vertical axis against the DGE levels in TPM on the horizontal axis. A strong inverse correlation ($r = -0.68$) between Δ Ct value and DGE expression is observed.

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.^{35,40,41} 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,^{11,42} we investigated

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

Subtype	IPA ^a	P-value ^b	Pathway-associated genes ^c (upregulated/downregulated)
dic(9;20)	TP53 signaling	0.032	<i>CDKN2A</i> (--), <i>CSNK1D</i> (+), <i>CTNBN1</i> (+), <i>LRDD</i> (+), <i>TP73</i> (+++), <i>JUND</i> (+), <i>JUNB</i> (+)
t(9;22)	Interferon signaling	0.000089	<i>IFIT1</i> (++), <i>IFIT3</i> (+), <i>IRF9</i> (+), <i>OAS1</i> (+), <i>STAT1</i> (+), <i>STAT2</i> (+)
	Activation of IRF by cytosolic pattern recognition receptors	0.00034	<i>DHX58</i> (+), <i>IFIT2</i> (+), <i>IRF9</i> (+), <i>NFKB2</i> (+), <i>NFKBE</i> (+), <i>STAT1</i> (+), <i>STAT2</i> (+)
t(12;21)	Role of RIG1-like receptors in antiviral innate immunity	0.002	<i>CASP8</i> (+), <i>CASP10</i> (+), <i>DHX58</i> (+), <i>NFKB2</i> (+), <i>NFKBE</i> (+)
	Anti-apoptosis	0.0026	<i>BAG3</i> (+), <i>BCL2L1</i> (+), <i>BIRC7</i> (++), <i>CD27</i> (+)
HeH	VDR/RXR activation	0.014	<i>KLF4</i> (+), <i>NCOR2</i> (+), <i>PDGFA</i> (-), <i>PRKCQ</i> (+)
	Apoptosis signaling	0.00081	<i>BCL2L1</i> (-), <i>CAPN2</i> (+), <i>CAPN3</i> (-), <i>CASP8</i> (-), <i>CASP10</i> (-), <i>NFKB2</i> (-), <i>NFKBIB</i> (-), <i>NFKBIE</i> (-), <i>PRKCQ</i> (-)
	OX40 signaling	0.004	<i>BCL2L1</i> (-), <i>HLA-DOB</i> (--), <i>NFKB2</i> (-), <i>NFKBIB</i> (-), <i>NFKBIE</i> (-), <i>TNFSF4</i> (-)

Abbreviations: ALL, acute lymphoblastic leukemia; IPA, ingenuity pathway analysis; IRF, interferon-regulatory factor; RIG1, retinoic acid inducible gene-1; VDR/RXR, vitamin D3 receptor/retinoid X receptor. ^aIPA canonical pathways or IPA tox lists. ^bFisher's exact test P-value. ^cGene 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.

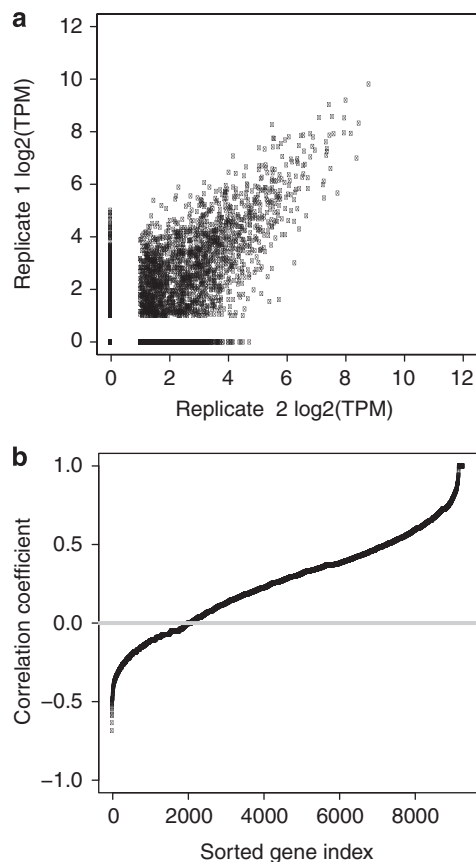


Figure 4. Detection of antisense transcription by digital gene expression (DGE). (a) Positive correlation between expression levels of antisense tags in DGE replication experiments from the same ALL patient RNA sample (Pearson's correlation coefficient $r = 0.70$). (b) Lack of systematic correlation 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 *NalIII* 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.²⁵ 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 ($P < 1 \times 10^{-15}$). 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.

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

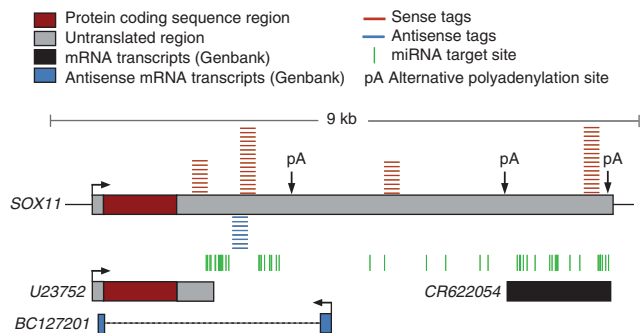


Figure 5. Schematic view of the sense and antisense tags in the *SOX11* gene region. The positions of the tags in the sense orientation of *SOX11* are indicated in red and those of the *SOX11*-antisense tag is indicated in blue, miRNA target sites (TargetScan26) are illustrated as vertical green lines and the poly(a) sites are indicated by pA (PolyA DB). This figure is not drawn to scale.

genes have been identified and breakpoints of the dic(9;20) aberration are heterogeneous, suggesting that the resulting loss of DNA at the 9p or 20q regions may have a functional role.^{44,45} The differentially regulated genes that we identified in dic(9;20) cells were enriched to the TP53 signaling pathway. In each of the three samples of the dic(9;20) subtype, we observed specific downregulation of *CDKN2A* on chr 9p21. *CDKN2A* encodes two tumor-suppressor proteins, namely p16^{INK4a} and p14^{ARF}, and is frequently deleted in BCP ALL.⁴⁶⁻⁴⁸ *SMAD1*, which also belongs to the TP53 pathway, was downregulated in parallel with *CDKN2A*; a similar correlation has been reported in ALL samples with 9p deletions,⁴⁹ but not specifically in samples with dic(9;20). *JUNB*, *JUND* and *TP73*, which inhibit TP53⁵⁰ were also overexpressed. In the dic(9;20) subtype, the antisense transcripts at the *LIN9* and *JUND* gene regions were upregulated and downregulated at the *PSMD10* region. The proteins encoded by *LIN9* and *JUND* interact with genes in the TP53 signaling pathway, and *PSMD10* is specifically involved in TP53 degradation,⁵¹ further pointing at the importance of this signaling pathway in ALL patients with the dic(9;20) subtype.

Different anti-apoptotic mechanisms in the HeH and t(12;21) subtypes were suggested by the genes highlighted in the pairwise analyses. The HeH subtype is characterized by > 50 chromosomes in leukemic cells, but whether the aneuploidy is a cause or a consequence of the leukemic transformation is unclear.⁵² In our HeH samples, seven genes involved in signaling pathways that regulate apoptosis were differentially expressed. Six genes were downregulated in HeH cells, including *BCL2L1* and *PRKCCQ* that arrest cell-cycle progression.^{53,54} As aneuploidy arises from aberrant control of mitotic checkpoints, we speculate that the lower expression of such genes may contribute to the HeH phenotype. In t(12;21) cells, DGE identified four upregulated genes in anti-apoptosis signaling, including *BCL2L1*, *BAG3*, *CD27* and *BIRC7*. Genes in the *BCL2* family can act as anti- or pro-apoptotic regulators, and the longer isoform of *BCL2L1* acts as an apoptosis inhibitor, whereas the shorter isoform acts as an apoptotic activator.⁵⁵ We could not distinguish between the isoforms because they share the same 3' UTR sequence; however, we observed the highest expression coming from the most 3' tag in the gene for the t(12;21) samples (average 51.1 TPM), indicating that the transcript with the longer 3' UTR is predominantly transcribed. The genes in the pathways highlighted in t(9;22) cells include *STAT1*, *STAT2*, *IRF9*, *NFKBE* and *NFKB2*, all of which are upregulated and belong to the JAK/STAT signaling cascade (Table 2).

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,⁵⁶ through DNA-RNA interactions, or through transcription interference.⁵⁷ For example, the tumor-suppressor gene *P15* has an antisense RNA that silences the *P15* gene through epigenetic alteration of heterochromatin in leukemia cell lines.¹⁸ 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,⁵⁸ and the Notch pathway is frequently activated in T-ALL.⁵⁹ 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.⁴¹ In our study, we also observed extensive APA and a strong association between genes with APA and antisense expression in the same gene region ($P < 1 \times 10^{-15}$). 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 Lindersson and Kristina Larsson. Affymetrix data were generated at the Uppsala Array Platform (<http://www.medsci.uu.se/klinfarm/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 Milani for assistance with the experimental design, Annabeth H Petersen for helpful discussions, Anders Lundmark for statistical expertise, Mårten Fryknäs for assistance with OncoPrint and Matilda Canderyd 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 (Rbc08-008). 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: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531-537.

- 2 Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz 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 Holleman A, Cheok MH, den Boer ML, Yang W, Veerman 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, Høgh AL, Emmersen 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, Shipitsin M, Maruyama R, Bessarabova M et al. Gene expression profiling of human breast tissue samples using SAGE-Seq. *Genome Res* 2010; **20**: 1730–1739.
- 10 t Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 2008; **36**: e141.
- 11 Morrissy AS, Morin RD, Delaney A, Zeng T, McDonald H, Jones S et al. Next-generation tag sequencing for cancer gene expression profiling. *Genome Res* 2009; **19**: 1825–1835.
- 12 The Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* 2009; **457**: 1028–1032.
- 13 Oszolaf F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 2010; **143**: 1018–1029.
- 14 He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The antisense transcriptomes of human cells. *Science* 2008; **322**: 1855–1857.
- 15 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**: 1559–1563.
- 16 Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M et al. Antisense transcription in the mammalian transcriptome. *Science* 2005; **309**: 1564–1566.
- 17 Ebralidze AK, Guibal FC, Steidl U, Zhang P, Lee S, Bartholdy B et al. PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. *Genes Dev* 2008; **22**: 2085–2092.
- 18 Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* 2008; **451**: 202–206.
- 19 Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 2008; **320**: 1643–1647.
- 20 Mayr C, Bartel DP. Widespread shortening of 3' UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 2009; **138**: 673–684.
- 21 Galante PA, Vidal DO, de Souza JE, Camargo AA, de Souza SJ. Sense-antisense pairs in mammals: functional and evolutionary considerations. *Genome Biol* 2007; **8**: R40.
- 22 Morrissy AS, Griffith M, Marra MA. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res* 2011; **21**: 1203–1212.
- 23 Milani L, Lundmark A, Kiialainen A, Nordlund J, Flaegstad T, Forestier E et al. DNA methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. *Blood* 2010; **115**: 1214–1225.
- 24 Zhang Y, Li J, Kong L, Gao G, Liu QR, Wei L. NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res* 2007; **35**: D156–D161.
- 25 Koscielny G, Le Texier V, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F et al. ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics* 2009; **93**: 213–220.
- 26 Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005; **120**: 15–20.
- 27 Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods* 2001; **25**: 402–408.
- 28 Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004; **5**: R80.
- 29 Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 2008; **9**: 321–332.
- 30 Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 2002; **99**: 6567–6572.
- 31 Witten D, Tibshirani R, Gu SG, Fire A, Lui WO. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol* 2010; **8**: 58.
- 32 Bishop JO, Morton JG, Rosbash M, Richardson M. Three abundance classes in HeLa cell messenger RNA. *Nature* 1974; **250**: 199–204.
- 33 Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18 000 cancer gene expression profiles. *Neoplasia* 2007; **9**: 166–180.
- 34 Ohno T, Kimura Y, Sakurada K, Sugimura K, Fujiyoshi T, Saheki T et al. Argininosuccinate synthetase gene expression in leukemias: potential diagnostic marker for blastic crisis of chronic myelocytic leukemia. *Leuk Res* 1992; **16**: 475–483.
- 35 Mozos A, Royo C, Hartmann E, De Jong D, Baro C, Valera A et al. SOX11 expression is highly specific for mantle cell lymphoma and identifies the cyclin D1-negative subtype. *Haematologica* 2009; **94**: 1555–1562.
- 36 Gold MR, Ingham RJ, McLeod SJ, Christian SL, Scheid MP, Duronio V et al. Targets of B-cell antigen receptor signaling: the phosphatidylinositol 3-kinase/Akt/glycogen synthase kinase-3 signaling pathway and the Rap1 GTPase. *Immunol Rev* 2000; **176**: 47–68.
- 37 Juric D, Lacayo NJ, Ramsey MC, Racevskis J, Wiernik PH, Rowe JM et al. Differential gene expression patterns and interaction networks in BCR-ABL-positive and -negative adult acute lymphoblastic leukemias. *J Clin Oncol* 2007; **25**: 1341–1349.
- 38 Morris KV. The emerging role of RNA in the regulation of gene transcription in human cells. *Semin Cell Dev Biol* 2011; **22**: 351–358.
- 39 Grinchuk OV, Jenjaroenpun P, Orlov YL, Zhou J, Kuznetsov VA. Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns. *Nucleic Acids Res* 2010; **38**: 534–547.
- 40 Wang X, Asplund AC, Porwit A, Flygare J, Smith CI, Christensson B et al. The subcellular Sox11 distribution pattern identifies subsets of mantle cell lymphoma: correlation to overall survival. *Br J Haematol* 2008; **143**: 248–252.
- 41 Gustavsson E, Sernbo S, Andersson E, Brennan DJ, Dictor M, Jerkeman M et al. SOX11 expression correlates to promoter methylation and regulates tumor growth in hematopoietic malignancies. *Mol Cancer* 2010; **9**: 187.
- 42 Maruyama R, Shipitsin 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.
- 43 Zachariadis V, Gauffin F, Kuchinskaya E, Heyman M, Schoumans J, Blennow E et al. The frequency and prognostic impact of dic(9;20)(p13.2;q11.2) in childhood B-cell precursor acute lymphoblastic leukemia: results from the NOPHO ALL-2000 trial. *Leukemia* 2011; **25**: 622–628.
- 44 Schoumans J, Johansson B, Corcoran M, Kuchinskaya E, Golovleva I, Grandt 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; **135**: 492–499.
- 45 Sulong S, Moorman AV, Irving JA, Strefford JC, Konn ZJ, Case MC et al. A comprehensive analysis of the CDKN2A gene in childhood acute lymphoblastic leukemia reveals genomic deletion, copy number neutral loss of heterozygosity, and association with specific cytogenetic subgroups. *Blood* 2009; **113**: 100–107.
- 46 Schiffman JD, Wang Y, McPherson LA, Welch K, Zhang N, Davis R et al. Molecular inversion probes reveal patterns of 9p21 deletion and copy number aberrations in childhood leukemia. *Cancer Genet Cytogenet* 2009; **193**: 9–18.
- 47 Forestier E, Gauffin F, Andersen MK, Autio K, Borgstrom G, Golovleva I et al. Clinical and cytogenetic features of pediatric dic(9;20)(p13.2;q11.2)-positive B-cell precursor acute lymphoblastic leukemias: a Nordic series of 24 cases and review of the literature. *Genes Chromosomes Cancer* 2008; **47**: 149–158.
- 48 Zhang Y, Xiong Y, Yarbrough WG. ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the Rb and p53 tumor suppression pathways. *Cell* 1998; **92**: 725–734.

- 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.



This work is licensed under the Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>

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