Supporting Information

Eisfeld et al. 10.1073/pnas.1203756109

SI Materials and Methods

The custom GeneAnnot Chip Description File (CDF) was used for gene expression value computation, which results in a single probe set for each (1). Summary measures of gene expression were computed for each probe set using the robust multichip average method, which incorporates quantile normalization of arrays. Expression values were logged (base 2) before analysis. The data were obtained for three different sets of younger patients enrolled in Cancer and Leuemia Group B (CALGB) 9621 (subgroup 1; $n = 53$) and CALGB 19808 (subgroup 2; $n = 90$) and older patients enrolled in CALGB 9720 and 10201 (subgroup 3; $n = 110$). To avoid batch effects, we performed separate analyses for each of the three subgroups. Patients were grouped into *BAALC* high/low expressers and *RUNX1* high/low expressers using the median for each batch as the cutoff.

For a subgroup of patients ($n = 118$), we confirmed the *BAALC* expression levels derived from the microarray data with quantitative real-time reverse-transcription PCR (RT-PCR). One microgram total RNA was reverse transcribed into cDNA using SuperScriptIII (Invitrogen) in a 40-μL reaction mix according to protocol instructions. The TaqMan assays were carried out for each sample in triplicate using Taqman Primer-Probe sets for *BAALC* and *ABL* (Life Technologies/Applied Biosystems) according to protocol instructions. To determine the relative levels of expression of *BAALC*, the comparative C_T method was used (Life Technologies/Applied Biosystems). First, the parameter threshold cycle (C_T) was determined for *BAALC* and *ABL*, and the cycle number difference (*ABL* – *BAALC* = ΔC_T) was calculated for each replicate. If *BAALC* failed to reach the software set threshold, the sample was considered below detection limit. If *ABL* amplification failed, the sample was omitted from the analysis. Finally, the mean ΔC_T from the three replicates was generated [$(\Sigma\Delta C_T)/3 = MC_T$], normalizing *BAALC* expression to *ABL* expression.

Correlation analyses using continuous data revealed a high reproducibility of *BAALC* expression levels with both techniques (subgroup 1: $R = 0.71$, subgroup 2: $R = 0.83$, subgroup 3: $R = 0.91$) (Fig. S2).

1. Ferrari F, et al. (2007) Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics* 8:446.

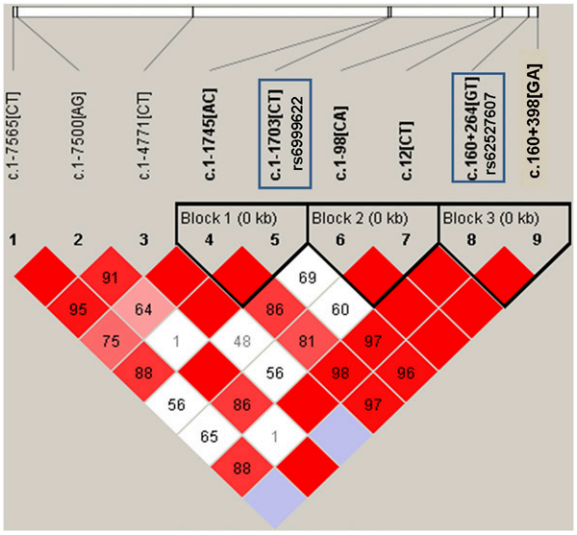

Fig. S1. Haplotype structure of *BAALC*, determined using genotype data of the nine marker SNPs from cases and controls. Haploview 4.2 (1) was used to obtain the linkage disequilibrium (LD) plot of nine SNPs using combined case and control samples, on the basis of measures of D′ and LOD values. Diamonds represent pairwise LD between SNPs, with darker shading of red indicating stronger LD (D′ ≥ 0.8 and LOD ≥ 2.0). Red boxes with no numbers indicate D′ = 1 and LOD ≥ 2. Blocks were defined on the basis of the confidence interval method (2). The two SNPs associated with *BAALC* expression: rs6999622 (c.1–1703[CT]) and rs62527607 (c.160+264[GT]) are in almost complete linkage disequilibrium with each other [LD (rs6999622, rs62527607) D′= 0.97, $R^2 = 0.92$, LOD = 136.35].

1. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.
2. Gabriel SB, et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.

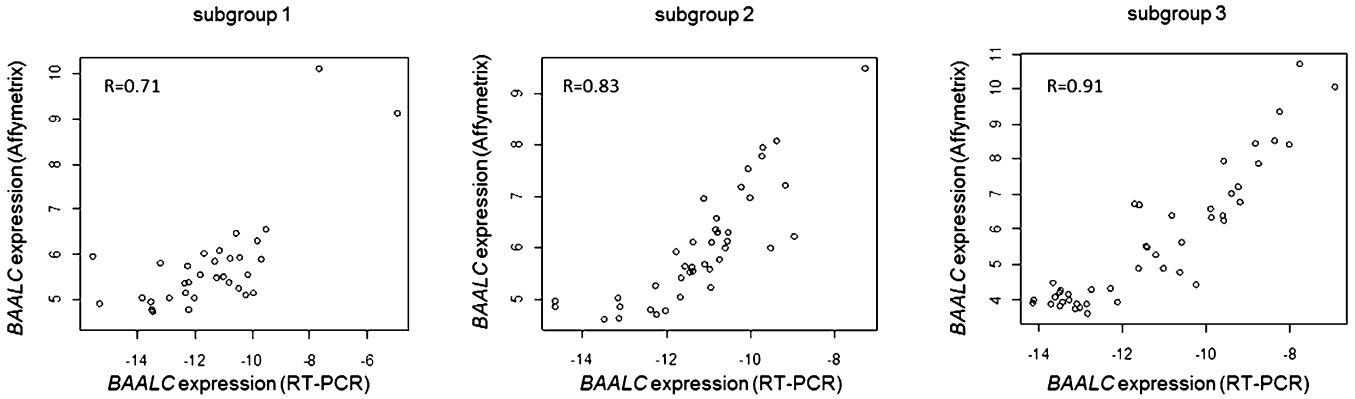

Fig. S2. Correlation of *BAALC* expression levels obtained via Affymetrix (using U133 plus 2.0 arrays) with *BAALC* expression levels obtained by quantitative real-time PCR (log *BAALC*/*ABL*) in three patient subgroups: younger patients enrolled in CALGB 9621 (subgroup 1; $n = 53$) and CALGB 19808 (subgroup 2; $n = 90$) and older patients enrolled in CALGB 9720 and 10201 (subgroup 3; $n = 110$).

Table S1. *BAALC* haplotypes observed in CN-AML patients and unaffected controls

Haplotype (number observed)	Cases (proportion observed)	Controls (proportion observed)
CACACCCGG (264)	129 (0.25)	135 (0.24)
CACACCCTA (2)	2 (0.004)	0
CACACCTGG (3)	1 (0.002)	2 (0.003)
CACACATGG (100)	50 (0.10)	50 (0.09)
CACATCCGG (2)	2 (0.004)	0
CACATCCTG (3)	2 (0.004)	1
CACATCCTA (120)	63 (0.12)	57 (0.10)
CACATATGG (4)	2 (0.04)	2 (0.04)
CACCCCCGG (263)	122 (0.24)	141 (0.25)
CACCCCCTA (1)	0	1 (0.002)
CACCCATGG (4)	0	4 (0.007)
CATATCCTG (2)	1 (0.001)	1 (0.002)
CGCACCCGG (4)	3 (0.006)	1 (0.002)
CGCATCCGG (1)	0	1 (0.002)
CGCCCCCGG (252)	111 (0.22)	141 (0.25)
CGCCCCTGG (2)	0	2 (0.004)
TGCACCCGG (1)	1 (0.002)	0
TGCCCCCGG (3)	0	3 (0.005)
TGTATCCGG (1)	1 (0.002)	0
TGTATCCTG (40)	16 (0.03)	24 (0.04)

Haplotypes were generated using the alleles of the nine SNP markers identified by sequencing and genotyping of cases ($n = 253$) and controls ($n = 286$, Fig. 1). Generation of haplotypes was performed using the PHASE v2.1.1 program. SNP markers are displayed in the table on the basis of their chromosomal location with the first SNP being the most upstream marker in relation to *BAALC* exon 1. No differences in haplotype proportions could be observed comparing cases and controls.

Table S2. *BAALC* haplotypes observed in CN-AML patients (*n* = 253) comparing high and low *BAALC* expressers

Haplotypes (number observed)	Low *BAALC* (proportion observed)	High *BAALC* (proportion observed)
CACACCCGG (129)	75 (0.27)	54 (0.23)
CACACCC**T**A (2)	2 (0.01)	0
CACACCTGG (1)	1 (0.004)	0
CACACATGG (50)	24 (0.09)	26 (0.11)
CACA**T**CCGG (2)	0	2 (0.01)
CACA**T**CC**T**G (2)	0	2 (0.01)
CACA**T**CC**T**A (63)	25 (0.09)	38 (0.16)
CACA**T**ATGG (2)	0	2 (0.01)
CACCCCCGG (122)	75 (0.28)	47 (0.2)
CATA**T**CC**T**G (1)	0	1 (0.004)
CGCACCCGG (3)	2 (0.01)	1 (0.004)
CGCCCCCGG (111)	59 (0.22)	52 (0.22)
TGCACCCGG (1)	0	1 (0.004)
TGTATCCGG (1)	0	1 (0.004)
TGTA**T**CC**T**G (16)	5 (0.02)	11 (0.05)

Haplotypes consist of the alleles of the nine SNP markers and were generated using the PHASE v2.1.1 program. The order of the SNP markers is based on their chromosomal location (Fig. 1). Two of the nine SNPs were associated with higher *BAALC* expression: rs6999622[CT] and rs62527607[GT]). Here rs6999622[CT] is listed at position 5 and rs62527607[GT] is listed at position 8 of the haplotypes (risk alleles are highlighted in bold). Patients harboring at least one copy of the T allele of either SNP were more likely to belong to the high *BAALC* expressing group ($P = 4.35 \times 10^{-4}$, risk haplotypes are shaded).

Table S3. Primer sequences and PCR conditions used for sequencing of the *BAALC* genomic region

Amplicon	Primer sequence (5′–3′)	Annealing temperature, °C
BAALC exon 1 F	TCCTGCCTCCCCAAATCAG	60
BAALC exon 1 R	AAAGTGCAGAACTAGCGATG	
BAALC exon 6 F	TTTGCAACATCTGCCATGTG	58
BAALC exon 6 R	CTACCCCCATTCTCCAGTC	
BAALC exon 8 F	GGTTTACATTTCCTAGTAACTC	58
BAALC exon 8 R	ACTGAACTGCACATTTGCAG	
BAALC intron 1 F	CACTGATCAGTGGACAGATG	60
BAALC intron 1 R	TGGGTGGCTGAGGGATATG	
BAALC promoter 1 F	GAGGGTTGAGATCATCTATG	56
BAALC promoter 1 R	AGACTGTGTGAACATGTTTC	
BAALC promoter 2 F	GAAAGCACAGGCTCTGCTAG	60
BAALC promoter 2 R	CAGTTGAGGGTAGAGCTGTC	
BAALC promoter 3 F	TCTACAGTGTTTAATTCCAC	58
BAALC promoter 3 R	CCAATTTCGTTCCCCACATC	