



## Genomic profiling of thousands of candidate polymorphisms predicts risk of relapse in 778 Danish and German childhood acute lymphoblastic leukemia patients

Wesoowska-Andersen, A; Borst, L; Dalgaard, M D; Yadav, R; Rasmussen, K K; Wehner, P S; Rasmussen, Morten; Ørntoft, T F; Nordentoft, I; Koehler, R; Bartram, C R; Schrappe, M; Sicheritz-Ponten, T; Gautier, L; Marquart, H; Madsen, H O; Brunak, S; Stanulla, M; Gupta, R; Schmiegelow, K.

*Published in:*  
Leukemia

*DOI:*  
[10.1038/leu.2014.205](https://doi.org/10.1038/leu.2014.205)

*Publication date:*  
2015

*Document version*  
Publisher's PDF, also known as Version of record

*Citation for published version (APA):*  
Wesoowska-Andersen, A., Borst, L., Dalgaard, M. D., Yadav, R., Rasmussen, K. K., Wehner, P. S., ... Schmiegelow, K. (2015). Genomic profiling of thousands of candidate polymorphisms predicts risk of relapse in 778 Danish and German childhood acute lymphoblastic leukemia patients. *Leukemia*, 29(2), 297-303. <https://doi.org/10.1038/leu.2014.205>

## ORIGINAL ARTICLE

# Genomic profiling of thousands of candidate polymorphisms predicts risk of relapse in 778 Danish and German childhood acute lymphoblastic leukemia patients

A Wesołowska-Andersen<sup>1,11</sup>, L Borst<sup>2,11</sup>, MD Dalgaard<sup>1</sup>, R Yadav<sup>1</sup>, KK Rasmussen<sup>2</sup>, PS Wehner<sup>3</sup>, M Rasmussen<sup>4</sup>, TF Ørntoft<sup>5</sup>, I Nordentoft<sup>5</sup>, R Koehler<sup>6</sup>, CR Bartram<sup>6</sup>, M Schrappe<sup>7</sup>, T Sicheritz-Ponten<sup>1</sup>, L Gautier<sup>1</sup>, H Marquart<sup>8</sup>, HO Madsen<sup>8</sup>, S Brunak<sup>1</sup>, M Stanulla<sup>9</sup>, R Gupta<sup>1,12</sup> and K Schmiegelow<sup>2,10,12</sup>

Childhood acute lymphoblastic leukemia survival approaches 90%. New strategies are needed to identify the 10–15% who evade cure. We applied targeted, sequencing-based genotyping of 25 000 to 34 000 preselected potentially clinically relevant single-nucleotide polymorphisms (SNPs) to identify host genome profiles associated with relapse risk in 352 patients from the Nordic ALL92/2000 protocols and 426 patients from the German Berlin–Frankfurt–Munster (BFM) ALL2000 protocol. Patients were enrolled between 1992 and 2008 (median follow-up: 7.6 years). Eleven cross-validated SNPs were significantly associated with risk of relapse across protocols. SNP and biologic pathway level analyses associated relapse risk with leukemia aggressiveness, glucocorticosteroid pharmacology/response and drug transport/metabolism pathways. Classification and regression tree analysis identified three distinct risk groups defined by end of induction residual leukemia, white blood cell count and variants in myeloperoxidase (*MPO*), estrogen receptor 1 (*ESR1*), lamin B1 (*LMNB1*) and matrix metalloproteinase-7 (*MMP7*) genes, ATP-binding cassette transporters and glucocorticosteroid transcription regulation pathways. Relapse rates ranged from 4% (95% confidence interval (CI): 1.6–6.3%) for the best group (72% of patients) to 76% (95% CI: 41–90%) for the worst group (5% of patients,  $P < 0.001$ ). Validation of these findings and similar approaches to identify SNPs associated with toxicities may allow future individualized relapse and toxicity risk-based treatments adaptation.

*Leukemia* (2015) 29, 297–303; doi:10.1038/leu.2014.205

## INTRODUCTION

Host genome variants affect the complex biology involved in pharmacokinetics/pharmacodynamics and need to be addressed to identify the critical factors determining treatment outcomes. Childhood acute lymphoblastic leukemia (ALL) has for several reasons been a model disease for such research owing to its frequency, well-described epidemiology, clinical characteristics and biologic profiles within cytogenetically defined subsets. In general, childhood ALL is very chemosensitive, and patients are almost uniformly treated within collaborative groups that stratify treatment according to known molecular aberrations in the leukemic clone, clinical characteristics and treatment response.<sup>1,2</sup> Most contemporary treatment protocols achieve 5-year event-free survival rates above 80%.<sup>3</sup> However, even with risk group-adapted treatment there is a wide diversity in cure rates, partly explainable by both host and cancer genomes.<sup>3,4</sup> Several candidate gene studies have associated inherited polymorphisms with treatment response and cure rates in childhood ALL.<sup>1,2,5–7</sup> Still, the biologic

relevance and interpretation of multiple single-nucleotide polymorphisms (SNPs) identified in recent genome-wide association studies (GWAS) remain unclear, as the functions of several of the most significant genes are unknown,<sup>8</sup> and effect sizes of single SNP associations are in any case extremely limited. To address some of these limitations and complement the GWAS approach, we applied a multiplexed targeted sequencing method allowing screening of 25 000–34 000 preselected SNPs within biologic domains potentially relevant to childhood ALL, allowing both single variants and multiple SNPs acting in the same pathways to be explored for association with relapse, and finally combining it with known clinical risk factors in a predictive profile.<sup>9</sup>

## MATERIALS AND METHODS

Danish patients were 1–15 years of age and diagnosed 1992–2008 with B-cell precursor ALL (BCP-ALL) or T-lineage ALL and treated according to the Nordic Society for Pediatric Hematology and Oncology (NOPHO) ALL92 ( $N = 227$ ) or NOPHO ALL2000 protocols ( $N = 268$ ) (Figure 1, Supplementary

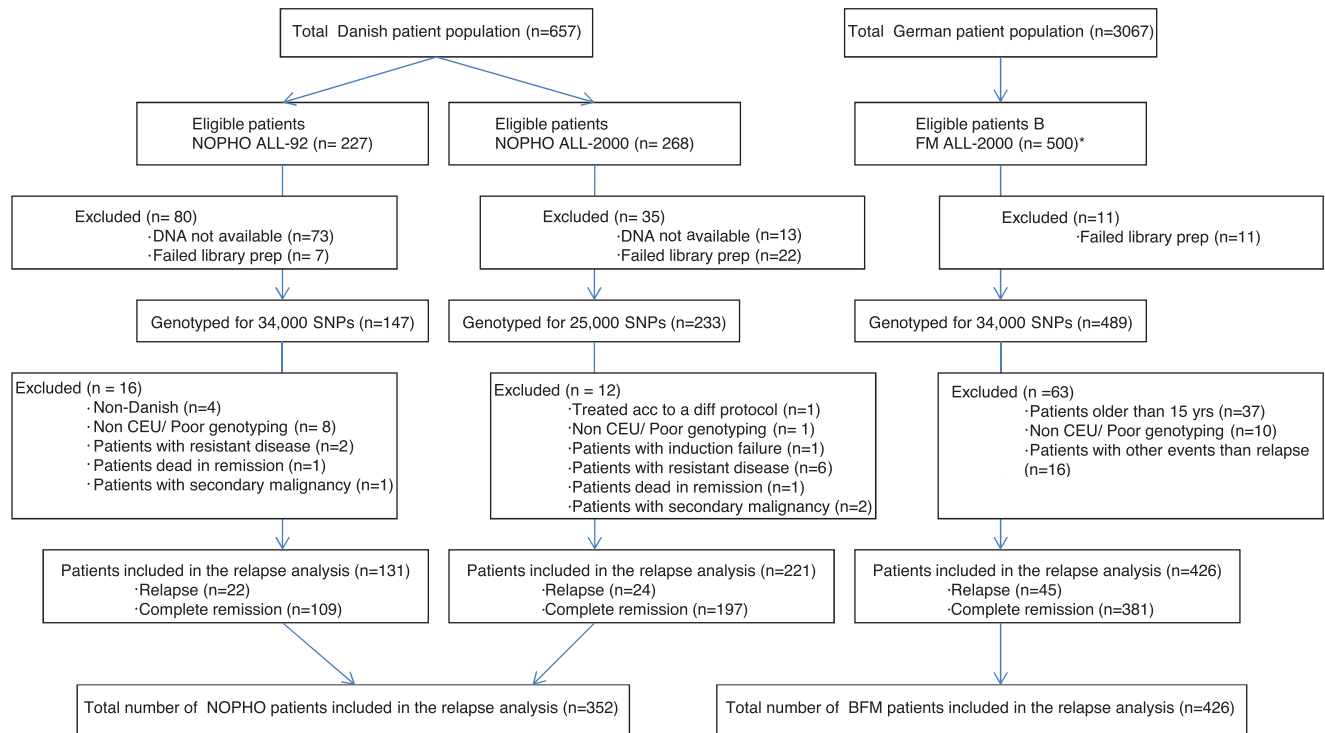
<sup>1</sup>Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark; <sup>2</sup>Pediatrics and Adolescent Medicine, The Juliane Marie Centre, The University Hospital Rigshospitalet, Copenhagen, Denmark; <sup>3</sup>Department of Pediatric Hematology and Oncology, HC Andersen Children's Hospital, Odense University Hospital, Odense, Denmark; <sup>4</sup>Centre for GeoGenetics, Natural History Museum of Denmark, The University of Copenhagen, Copenhagen, Denmark; <sup>5</sup>Institute of Clinical Medicine, Århus University Hospital, Århus, Denmark; <sup>6</sup>Department of Human Genetics, University of Heidelberg, Heidelberg, Germany; <sup>7</sup>Department of General Pediatrics, University Medical Center Schleswig-Holstein, Kiel, Germany; <sup>8</sup>Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany; <sup>9</sup>Department of Clinical Immunology, Diagnostic Centre, The University Hospital Rigshospitalet, Copenhagen, Denmark and <sup>10</sup>Institute of Clinical Medicine, Faculty of Health and Medical Sciences, The University of Copenhagen, Copenhagen, Denmark. Correspondence: Professor K Schmiegelow, Pediatrics and Adolescent Medicine, The Juliane Marie Centre, The University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark.

E-mail: kschmiegelow@rh.dk

<sup>11</sup>Joint first authorship.

<sup>12</sup>Joint senior authorship.

Received 16 April 2014; revised 14 June 2014; accepted 17 June 2014; accepted article preview online 3 July 2014; advance online publication, 25 July 2014



**Figure 1.** Patient flow. Overview of the patients included in the study. CEU, Utah residents with ancestry from northern and western Europe. \*BFM patients were selected upon availability of germline DNA and consecutively enrolled up to the predefined number of 500.

Online Material and Supplementary Tables 1). These two treatment protocols are very similar and included a 4-week, 3-drug induction phase (prednisolone, vincristine and doxorubicin (with intrathecal methotrexate)) with no glucocorticosteroid prephase, a risk group-adapted consolidation phase and methotrexate/6-mercaptopurine maintenance therapy up until 2 to 2.5 years from diagnosis<sup>10</sup> (detailed in Supplementary Online Material). End of induction minimal residual disease (MRD) measurements were available for 73% of NOPHO ALL2000 patients, but not included in risk stratification.<sup>10</sup> For cross-protocol validation of relapse-predictive host genomic variants, 500 German childhood ALL patients 1–18 years at diagnosis were included, all treated according to the BFM ALL2000 protocol. The BFM ALL2000 protocol included a 1-week prednisolone prephase, a 4-week, 4-drug induction phase (prednisolone or dexamethasone, vincristine, doxorubicin and L-asparaginase (with intrathecal methotrexate)), a risk group-adapted consolidation phase and methotrexate/6-mercaptopurine maintenance therapy until 2 years from diagnosis<sup>11,12</sup> (Supplementary Online Material). A total of 352 Danish and 426 German patients were eligible for the final relapse-risk analysis (Figure 1 and Supplementary Tables S1). Patients were enrolled between 1992 and 2008 with a median follow-up of 7.6 years (50% range: 5.3–9.2 years) for patients in first remission. In this study, both bone marrow and extramedullary relapses were events of interest. Owing to the complexity of the bioinformatic analysis and the application of nonlinear machine learning, and as it is uncertain how to weight competing events (induction failures, deaths in remission and second cancers), we chose, in this exploratory, hypothesis-generating study, to exclude patients with such events already at the time of diagnosis of ALL (Figure 1). The study was conducted in accordance with the Declaration of Helsinki Principles of 1975, and approved by the Danish Data Protection Agency, the Committee on Biomedical Research Ethics and by the Ethics Committee of the Hannover Medical School, Hannover, Germany.

SNP selection and bait design have been previously described in detail (Wesolowska *et al.*<sup>9</sup> and Supplementary Online Material). SNPs were selected to cover all known and putative clinically relevant genetic variation with regard to childhood ALL treatment (13 drugs), drug pharmacokinetics and pharmacodynamics, relapse risk and several toxicities (Supplementary Figure S1). In short, clinically important genes and SNPs for the 13 most commonly administered antileukemic drugs were evaluated with regard to genes encoding proteins involved in metabolism, transport, target proteins, regulation of drug-target response

and to some extent drug-related toxicity (e.g. coagulation, immune function) (Supplementary Online Material). Initially, this was carried out by literature curation, and then expanded to cover additional aspects of response to chemotherapy such as genes encoding proteins involved in apoptosis pathways and DNA repair. This also included consultations with experts within such areas and through various online resources (Supplementary Online Material). Finally, known drug–protein associations and first-order protein–protein interactions were evaluated and added (Supplementary Online Material and Supplementary Figure S1).<sup>9</sup> Approximately half of the selected genomic variants were in noncoding regions. The selected genes were then screened for known polymorphisms with putative functional consequence on their transcript (Supplementary Table S3). Baits for the SureSelect Target Enrichment System (protocol version 1.2 April 2009; Agilent Technologies, Santa Clara, CA, USA) were designed for all identified SNPs. The first design included baits targeting 25 086 clinically relevant SNPs in 1540 genes (Supplementary Figure S2). As baits covered more genomic material than just the targeted SNPs, readouts on the targeted regions covered 116 646 known human variants (Supplementary Table S4). This panel was subsequently updated after sequencing the first group of patients (from NOPHO ALL2000) (Supplementary Online Material and Supplementary Tables S3 and S4). Thus, the NOPHO ALL92 and BFM ALL2000 cohorts were screened using an expanded bait design, covering 33 683 SNPs in 2254 genes. The genomic targets of the two panels are shown in Supplementary Figure S2.

The patients were genotyped for either of the two SNP panels by multiplex targeted sequencing as described previously.<sup>9</sup> Briefly, the Agilent SureSelect Target Enrichment System protocol (protocol version 1.2 April 2009; Agilent Technologies) was modified, allowing nucleotide barcoding, and then pooling of samples before target enrichment and sequencing. Genomic DNA from leukemia remission samples was sheared, purified, end-repaired and 3' adenylated. Custom-made adaptors containing unique four-base barcodes were ligated to the DNA fragments before size selection and amplification. The DNA libraries were subsequently mixed in groups of up to eight samples, and the pooled libraries were hybridized to the custom-designed baits. Posthybridization amplification PCR was performed, and the pooled libraries were sequenced (Supplementary Online Material). The sequencing reads were mapped to the reference human genome (GRCh37), and SNP calling was performed (Supplementary Figure S3). The threshold set for SNP calling was minimum 10× sequencing depth. Data quality control and the individual–SNP association

analyses were performed in PLINK (version 1.07).<sup>13</sup> Patient samples not clustering together with the HapMap central European ancestry samples in the principal component analysis (Supplementary Figures S4 and S5) were removed.<sup>14</sup> To make sure that the data did not contain any bias arising either from sequencing or data processing, the observed minor allele frequencies (MAFs) of the genotyped SNPs were plotted against the HapMap CEU-reported MAFs for the 5962 SNPs, where data were available (Supplementary Figure S7). Furthermore, SNPs for which <50% of the patients could be genotyped at the minimum depth of 10× and SNPs with MAF <1% were excluded from the analyses. To validate readouts, genotyping from this experiment was compared with genotype calls obtained from the Illumina Human 1M-Duo SNP chip performed on a subset of patients (*N*=275, overlap of 2394 SNPs on both platforms). The mean genotyping concordance of those two methods was 98.2% (Supplementary Online Material).

### SNP associations

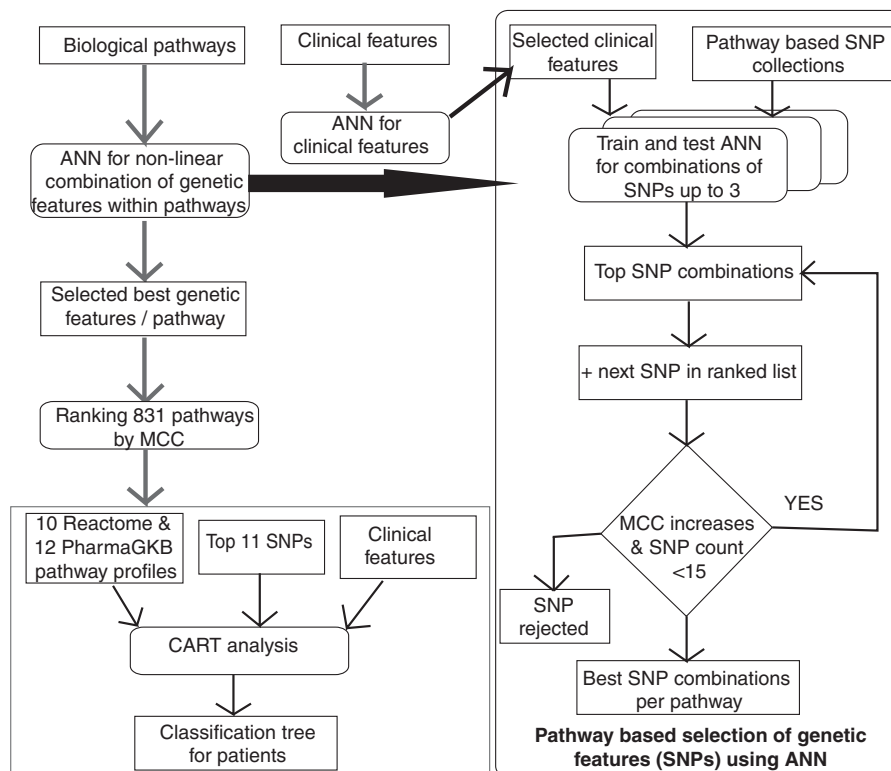
Single SNPs were explored for associations with risk of relapse using a Fisher's exact test separately for the NOPHO and BFM cohorts. *P*-values obtained were corrected for multiple testing by adaptive permutation, and subsequently only SNPs with adjusted *P*-values below 0.05 in both cohorts were considered significant.<sup>15</sup> Besides investigating such contribution of individual variations, the combined effects of multiple SNPs acting in the same biologic pathways were investigated using nonlinear machine learning. All pathways from the Reactome database<sup>16</sup> and 12-drug metabolism pathways from the PharmGKB database (ref. 27 in Supplementary Online Material), with up to 193 SNPs in a pathway, were

interrogated. Briefly, ANN models including all combinations of up to three SNPs together with white blood cell counts (WBC) and age at diagnosis were tested within each pathway and stepwise expanded with up to 15 SNPs in a pathway, if threefold cross-validated performance improved (Supplementary Online Material) (Figure 2). The best model for each pathway was then chosen, and the most informative pathway models were combined with clinical information in classification and regression tree (CART) analysis assigning for each pathway a score from 1 to 10 for every patient based on the relapse prediction from the neural network. As induction therapy differed between the Danish and German cohorts, each of these was dichotomously subdivided based on the approximate median end of induction MRD levels in each cohort being 10<sup>-3</sup> and 10<sup>-4</sup> in NOPHO and BFM cohorts, respectively (Supplementary Online Material).

### RESULTS

Of the genotyped patients (*N*=869), 808 fulfilled the quality control and the European ancestry criteria (Figure 1), of which 778 patients could be included in the final relapse-risk analysis. The majority of genotyped variants had very low heterogeneity (MAF <1%) between patients and thus did not contribute with sufficient power to the single SNP analysis. Thus, a total of 4260 and 3865 SNPs in the NOPHO and the BFM cohorts were included in the relapse-risk analysis, respectively.

The QQ plots showed good agreement with the null distribution and absence of genomic inflation (Supplementary Figure S8),



**Figure 2.** Flow diagram of the ANN models and CART (see also Supplementary Online Material). For pathway analysis, we included all nonsynonymous coding, frameshift coding, stop codon and splice site SNPs genotyped in this study with MAF above 0.005 residing in the pathway genes for pathways in Reactome database and for the 12-drug metabolism pathways from the PharmGKB database. Each pathway had between 1 and 193 SNPs, and each SNP was encoded by three values between 0 and 1 corresponding to likelihood of each genotype calculated from VCF file produced by SAMtools (see Supplementary Online Material). Associations with relapse risk were performed by training feedforward ANNs with backpropagation on subsets of SNPs from each pathway with threefold cross-validation. For each pathway, all combinations of up to three SNPs were assessed by means of MCC. The combinations were then further iteratively increased up to 15 SNPs by adding another SNP to the top 20 previous combinations of SNPs, if the MCC increased by at least 0.01. Pathways were then ranked by MCC of the best combination of SNPs for each pathway, and the most predictive pathways for relapse were then included in the CART analysis. This included the 426 patients with complete information on sex, age and WBC at diagnosis, immunophenotype, karyotype, end of induction MRD and risk group. For the large group of patients with low MRD, the SNP profiles of the top Reactome/PharmGKB pathways were included to explore their relapse prediction for this patient subset.

and several loci were significantly associated with risk of relapse (Supplementary Figure S9). A total of 188 and 152 SNPs were associated with relapse risk in the NOPHO and BFM cohorts, respectively (Supplementary Tables S5 and S6). As some of these will reflect chance findings, we included in the subsequent analyses only the 11 SNPs that were associated with relapse risk in both cohorts. Importantly, these 11 SNPs were related to risk of relapse independent of other known risk factors (Table 1, Supplementary Table S7 and Supplementary Online Material). As further support of their biologic significance, Kaplan–Meier analyses for these SNPs showed a general tendency of gene dose effects (0 versus 1 versus 2 alleles associated with increased risk of relapse) with log-rank trend *P*-values ranging from  $4.8 \times 10^{-6}$  for rs3216144 (matrix metalloproteinase 7, *MMP7*) to 0.03 for rs35721373 (dysferlin, *DYSF*), respectively (Supplementary Figure S9).

Next, functional SNPs were grouped by biologic pathways, and the relevance of the pathways to risk of relapse was assessed by training ANNs on different combinations of SNPs from each pathway, allowing nonlinear correlations between SNPs. The top-associated Reactome pathway<sup>16</sup> ranked by Matthew's correlation coefficient<sup>17</sup> (MCC, ranging from 0 to 1) was 'ATP-binding cassette (ABC) family protein-mediated transport' (MCC=0.33, area under the receiver-operator curve (AUC)=0.69) (Supplementary Table S8). Similarly, 12-drug metabolism pathways from the PharmGKB database relevant for the administered drugs were investigated,<sup>18</sup> and the top pathways were 'Vinka Alkaloid Pathway, Pharmacokinetics' (MCC=0.321, AUC=0.72), 'Glucocorticoid Pathway (Peripheral Tissue), Pharmacodynamics' (MCC=0.320, AUC=0.75) and 'Methotrexate Pathway (Brain Cell), Pharmacokinetics' (MCC=0.320, AUC=0.70) (Supplementary Table S9).

To explore and illustrate the combined significance of clinical data and host genomic findings, CART analysis was applied to subclassify sequentially patients in a multivariate model, including the 426 patients from both cohorts with complete information on sex, age, immunophenotype, WBC, leukemia karyotype, end of induction MRD, risk group and genotypes of the 11 cross-cohort-associated SNPs (Figure 3a). A total of 426 patients were included. End of induction MRD levels divided the patients into two major groups: one group with high MRD levels (i.e. above median for that cohort) and a high risk of relapse, which could be further stratified by SNPs in the myeloperoxidase (*MPO*), estrogen receptor 1 (*ESR1*), lamin B1 (*LMNB1*) and *MMP7* genes, and another large group with low MRD levels and low cumulative relapse risk. As the latter group, because of its size, accounts for

25% of all relapses, a subsequent CART analysis with pathway profiles for the top 10 Reactome and the PharmGKB pathways was added (all having AUCs of approximately 0.70), which indicated a role of ABC transporters and glucocorticosteroid transcription regulation pathways (Figure 3b). Finally, based on MRD, WBC, the aforementioned host genomic data and the resulting observed incidence of relapse (Figures 3a and b), we could define three large subsets of patients with significant differences in the risk of relapse ( $P < 0.001$ ). With 92% of projected relapses by Kaplan–Meier analysis having been observed in the total cohort, the two extreme subgroups of patients had 6-year cumulative risks of relapse of 4% (95% confidence interval: 1.6–6.3%) for the best outcome group (71.5% of all patients) and 76% (95% confidence interval: 41–90%) for the worst outcome group (5% of all patients) (Figure 3c and Table 2), leaving age, immunophenotype (BCP versus T-ALL), cytogenetics (Table 2; high risk, that is, t(9;22), hypodiploid, t(4;11); low risk, that is, high hyperdiploid, t(12;21); other) and risk group nonsignificant with this approach. As MRD was not available for all patients, an additional CART analysis was performed not including MRD (Supplementary Figure S10). All four SNPs from the original CART diagram remained significant for relapse prediction. Survival analysis taking into account the time to event was also performed (Supplementary Figure S11), with results largely agreeing with the model in Figure 3. Multivariate regression analysis to predict risk of relapse found all the features selected by CART analysis to be statistically significant, with MRD being the most significant clinical factor ( $P = 7.9 \times 10^{-6}$ ), *MPO* rs28730837 being the most significant SNP region ( $P = 0.002$ ) and glucocorticosteroid pathway (transcription regulation, pharmacodynamics) ( $P = 2.0 \times 10^{-13}$ ) being the most significant pathway.

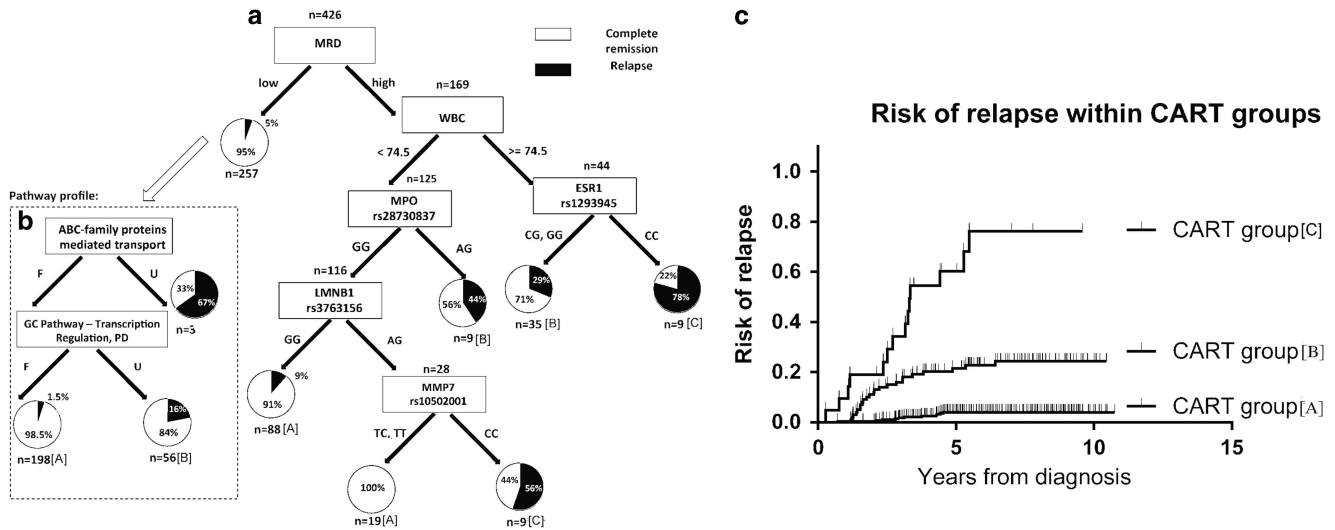
## DISCUSSION

Today, many high-risk patients do not respond to intensified treatment, but most relapses occur among non-high-risk patients. The present study, using genomic candidate gene genotyping and front-line bioinformatics analyses, provides a novel biology/pharmacology-driven approach for outcome prediction and goes beyond conventional genome-wide association studies (GWAS) approaches, while still being more cost-effective than whole-genome sequencing (for cost details see Supplementary Online Material). The rapidly growing understanding of the complex human genome and its derived functional biology allows selection of candidate SNPs based on the current knowledge of pharmacogenomics, disease mechanisms, signaling pathways and protein

**Table 1.** SNPs associated with risk of relapse discovered in both Danish and German cohorts

SNP			NOPHO ALL92 and 2000				BFM ALL2000				Combined cohorts			
Rs ID	Gene	Cons	P-value	OR	MAF relapse	MAF CR	P-value	OR	MAF relapse	MAF CR	MAF relapse	MAF CR	P-value	OR
Rs3216144	<i>MMP7</i>	Regulatory	4.0E-04	0.14	0.05	0.26	2.0E-03	0.32	0.09	0.24	0.07	0.24	6.0E-06	0.26
Rs10502001	<i>MMP7</i>	NSC	4.0E-04	0.14	0.05	0.25	2.0E-03	0.32	0.09	0.24	0.08	0.24	6.0E-06	0.26
Rs10795242	<i>AKR1C3</i>	Intronic	8.6E-03	2.13	0.31	0.17	3.7E-02	1.85	0.24	0.15	0.27	0.16	5.0E-04	2.01
Rs28730837	<i>MPO</i>	NSC	4.7E-02	4.26	0.06	0.02	1.5E-02	3.19	0.06	0.02	0.06	0.02	1.0E-03	3.60
Rs6139873	<i>CHGB</i>	NSC	4.0E-03	19.56	0.11	0.01	3.9E-02	3.50	0.04	0.01	0.06	0.01	1.1E-03	5.75
Rs1293945	<i>ESR1</i>	Regulatory	2.0E-02	1.86	0.59	0.43	2.6E-02	1.95	0.59	0.42	0.59	0.43	1.3E-03	1.91
Rs3763156	<i>LMNB1</i>	Intronic	2.0E-02	3.06	0.19	0.07	4.0E-02	2.15	0.15	0.07	0.16	0.07	1.6E-03	2.43
Rs55684978	<i>HTR3D</i>	SC	1.6E-02	8.08	0.07	0.01	3.2E-02	3.85	0.04	0.01	0.05	0.01	1.8E-03	4.88
Rs1058047	<i>TMED7</i>	Splice site	1.3E-02	8.64	0.08	0.01	3.9E-02	3.20	0.05	0.02	0.06	0.02	3.1E-03	3.98
Rs35721373	<i>DYSF</i>	SC	1.5E-02	4.89	0.13	0.03	3.1E-02	2.52	0.09	0.04	0.10	0.04	3.3E-03	2.96
Rs6601899	<i>AKR1C3</i>	Intronic	3.9E-02	1.77	0.31	0.20	3.6E-02	1.80	0.24	0.15	0.27	0.17	3.3E-03	1.84

Abbreviations: cons, consequence of the SNP on its transcript from Ensembl Variant Effect Predictor; MAF CR, minor allele frequency in complete remission patients; MAF relapse, minor allele frequency in relapse patients; NSC, nonsynonymous coding; OR, odds ratio; *P*-value, adaptive permutation *P*-values; SC, synonymous coding.



**Figure 3.** CART analysis of sequentially subclassified patients by clinical data including WBC, end of induction MRD and genotypes of cross-cohort relapse-associated SNPs for the 426 patients from both cohorts for whom these data were available. The most discriminatory WBC value ( $74.5 \times 10^9/l$ ) is selected by the CART algorithm. Black and white color in the pie charts represents the percentage of patients who experienced a relapse (black) or stayed in complete remission (white). One group with above median (for that cohort) MRD levels and a high risk of relapse could be further stratified by SNPs in the myeloperoxidase (*MPO*), estrogen receptor 1 (*ESR1*), lamin B1 (*LMNB1*) and *MMP7* genes (a). Another group with low MRD and low cumulative relapse risk could be further stratified by pathway profiles of ABC transporters and glucocorticosteroid transcription regulation pathways (b). (c) Kaplan–Meier plots of relapse risk for three subsets of patients identified by the CART analysis. The groups were defined by the observed incidence of relapse within each node of the graph in panels a and b as marked with [A], [B] or [C] for the best, intermediate and worst outcome group, respectively ( $P < 0.001$ ). Vertical lines depict patients with relapse or lack of further follow-up. F, favorable; GC, glucocorticosteroids; PD, pharmacodynamics; U, unfavorable.

interactions. Large-scale, genomic candidate gene setup facilitates not only single SNP investigations but also associations of multiple variations grouped by their putative function. Further associations of combinations of SNPs grouped by biologic pathways tested with neural network models enable detection of meaningful nonlinear SNP interactions. The results obtained through these strategies can provide conclusions at new levels of genomic complexity, collectively emphasizing the importance of specific biologic mechanisms for the phenotype.

The 11 cross-cohort relapse-associated SNPs resided in genes previously suggested as markers for leukemia aggressiveness, involved in steroid response, implicated in resistance mechanisms or toxicities of certain drugs (Supplementary Online Material). The pathway analysis strongly indicated an importance of the ABC family protein-mediated transport, activation of matrix metalloproteinases, toll-like receptor cascade signaling and various signal-transduction pathways, as well as involvement of the cell cycle. The top-associated pathway ‘ABC family protein-mediated transport’ confirms the role of pharmacogenomics in drug response by pointing to the importance of drug transport. The ATP-dependent drug efflux pumps have broad substrate specificities; they influence drug accumulation and are associated with the development of resistance to anticancer drugs.<sup>19</sup> Matrix metalloproteinases are involved in tumor progression,<sup>20</sup> whereas the toll-like receptor cascades and nucleotide-binding oligomerization domain-containing protein 1 and 2 signaling pathways support a role for the innate immune system function for proper drug response or ALL biology.<sup>21,22</sup>

It is noteworthy that the three top drug metabolism pathways for relapse risk corresponded to pharmacodynamics of glucocorticosteroids and pharmacokinetics of vinka alkaloids and methotrexate. This emphasizes the clinical significance of glucocorticosteroid therapy and cell cycle-arresting agents such as vincristine and methotrexate, and also highlights the potential profound consequence of adverse drug disposition on the effect of these agents.

As risk factors linked to host genomics, leukemia biology, treatment response markers and drug pharmacokinetics accumulate, CART analysis facilitates integration of the different layers of molecular complexity with patients’ clinical characteristics to identify groups of patients with distinct treatment outcomes, and accordingly candidates for such treatment adaption. Thus, combining patients’ genotypes with their clinical features is likely to explain treatment outcome better than single SNPs.

Even though the binary recursive partitioning of the CART methodology offers an approach to patient classifications that is easier to perceive from a biologic point of view and to apply in the clinical setting, CART does not necessarily outperform conventional, stepwise, multivariate regression analyses. Interactions between variables (and combinations hereof) may be missed, and the downstream decision branching may be unstable if one variable is removed. Although the CART analysis presented here classified patients into highly different prognostic subsets, and conventional regression analysis confirmed the most significant clinical feature, SNP and pathway, larger independent data sets will be needed to determine which of these two biostatistical approaches is superior for relapse prediction.

In future trials, the small group of patients who have the highest risk of relapse can be offered intensified treatments, or be candidates for phase 2 trials, whereas treatment reduction to avoid specific toxicities may be relevant for the large subset of patients who have a projected risk of relapse of  $< 5\%$ . In this context, the least relapse-predictive drug metabolism pathways may point to which drug doses potentially could be safely reduced in complex combination chemotherapy, and/or to which patients a specific drug is important owing to the patient’s genomic variants.

Although further verification is needed to strengthen the findings, the present approach has, when compared with other large-scale studies, the clear advantage of specifically targeting potential causative variants, which reduces the need to investigate genomic patterns of linkage disequilibrium, or conducting

**Table 2.** Clinical characteristics of the patients in the three CART groups

Characteristics	Group [A] patients	Group [B] patients	Group [C] patients
Total patients (%)	305 (71.5)	100 (23.5)	21 (5)
<i>Gender</i>			
Male	174 (57)	63 (63)	12 (57)
Female	131 (43)	37 (37)	9 (43)
Age (years), mean (50% range)	5.32 (2.89–6.71)	7.38 (3.54–10.97)	7.09 (3.75–10.88)
WBC × 10 <sup>9</sup> /l, mean (50% range)	22.88 (4.6–26.5)	100.36 (7.5–108)	90.61 (4.51–121)
<i>Immunophenotype</i>			
BCP-ALL	279 (91.5)	77 (77)	17 (81)
T-ALL	25 (8.2)	23 (23)	4 (19)
Other	1 (0.3)	0 (0)	0 (0)
<i>MRD</i>			
Low	237 (77.7)	67 (67)	12 (57)
High	68 (22.3)	33 (33)	9 (43)
<i>Cytogenetics</i>			
t(9;22)	4 (1.3)	1 (1)	1 (4.8)
t(1;19)	3 (1)	1 (1)	0 (0)
t(12;21)	79 (25.9)	13 (13)	2 (9.5)
t(4;11)	2 (0.65)	1 (1)	0 (0)
Hypodiploid	2 (0.65)	3 (3)	0 (0)
Hyperdiploid	81 (26.5)	16 (16)	1 (4.8)
Other	17 (5.6)	5 (5)	2 (9.5)
Normal/no data	120 (39.3)	61 (61)	15 (71.4)
<i>Relapse distribution</i>			
Total	11	23	14
Bone marrow	8 (72.7)	19 (82.6)	12 (85.7)
CNS	1 (9.1)	3 (13.1)	4 (28.6)
Other	2 (18.2)	2 (8.7)	0 (0)

Abbreviations: ALL, acute lymphoblastic leukemia; BCP-ALL, B-cell precursor ALL; CART, classification and regression tree; CNS, central nervous system relapses (isolated and combined CNS relapses); MRD, minimal residual disease; T-ALL, T-lineage acute lymphoblastic leukemia; WBC, white blood cell count.

follow-up fine-mapping studies. Functional prioritization of the potential SNPs to be genotyped eliminates the difficulties linked to GWAS when mapping significant variants to genes and genotypes, and can be more directly combined in functional analyses. As an example, only one of the 134 SNPs associated to relapse risk in the recent childhood ALL GWAS study of Yang *et al.*<sup>8</sup> was also genotyped in the present study, as the remaining SNPs resided in noncoding or functionally unannotated regions, and it was not found significant. Moreover, as the associated SNPs in most cases were difficult to map to the gene on which they exert their effects, it is difficult to evaluate the involved biologic mechanisms.

The present study combines large-scale genetic investigations of a GWAS approach with the targeted focus of a candidate gene approach. Thus, the major strength of our approach is that a wide panel of functional SNPs was selected with various prior assumptions of their potential relevance for childhood ALL treatment outcome. Owing to annotation to specific pathways, the results obtained are easier to interpret, and eventually to implement in the clinic based on existing knowledge of ALL disease mechanisms and pharmacokinetics and pharmacodynamics of administered drugs. Thus, the SNP/pathway-based approach can also be applied for prediction of specific side effects of antileukemic therapy. Host genome profiles may then indicate drugs with little importance for cure but associated with a high risk for side effects for specific patients, and this information could be applied for tailored therapy. Accordingly, the strategies and findings of this paper will be validated in the ongoing Nordic/

Baltic NOPHO ALL2008 protocol addressing risk grouping,<sup>23</sup> occurrence of toxicities<sup>24</sup> (including infections<sup>25</sup>) and risk of relapse before they are included into future treatment allocation.

We acknowledge that not covering the whole genome will certainly miss unknown, important genetic components of treatment response, which may benefit from alternative genome-wide screening approaches such as GWAS or whole-exome/transcriptome/genome sequencing, although the latter is still burdened by high costs owing to required patient numbers for statistical significance, and requires extensive follow-up for unannotated variants to be credibly accepted.

In conclusion, this large-scale integration of the knowledge of disease mechanisms and drug pharmacokinetics/pharmacodynamics in host genome studies in childhood ALL can offer both significant improvements to the current relapse prediction algorithms, and, importantly, indicate specific directions for leukemia response and toxicity risk-based treatments adaptation for the individual patient.

#### CONFLICT OF INTEREST

The authors declare no conflict of interest.

#### ACKNOWLEDGEMENTS

This study was supported by grants from the Danish Cancer Society (Grant numbers R2-A56-09-S2 and R20-A1156-10-S2), the Danish Childhood Cancer Foundation, the Ministry of Health (Grant number 2006-12103-250), the Novo Nordisk Foundation and the Danish Research Council for Health and Disease (Grant numbers 271-06-0278,

271-08-0684). KS holds the Danish Childhood Cancer Foundation Professorship in Paediatric Oncology. RG is supported by a grant from the Danish National Research Foundation to the Sino-Danish Breast Cancer Research Center. The study was conducted in the framework of the International BFM Study Group.

## AUTHOR CONTRIBUTIONS

KS had full access to all of the data in the study and takes responsibility for the integrity of the data, and the accuracy of the data analysis. KS and RG conceived and headed the project. AWA and LB wrote the manuscript. AWA performed the bioinformatic analyses and compiled the work. LB, MDD and KKR performed the laboratory work. TSP and RG assisted with the neural network analysis. MSc and MSt provided samples from BFM cohort. RK and CRB performed MRD analyses on the German cohort. PSW provided samples from western Denmark. MR, EW, TØ and IN performed the sequencing. LG provided technology consultation and laboratory resources. SB provided computational infrastructure. All authors provided critical input to the project and manuscript and approved the final manuscript. Professor Eske Willerslev is thanked for assistance with genome sequencing.

## REFERENCES

- Stanulla M, Schaeffeler E, Flohr T, Cario G, Schrauder A, Zimmermann M *et al*. Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia. *JAMA* 2005; **293**: 1485–1489.
- Davidson ML, Dalhoff K, Schmiegelow K. Pharmacogenetics influence treatment efficacy in childhood acute lymphoblastic leukemia. *J Pediatr Hematol Oncol* 2008; **30**: 831–849.
- Pui CH, Mullighan CG, Evans WE, Relling MV. Pediatric acute lymphoblastic leukemia: where are we going and how do we get there? *Blood* 2012; **120**: 1165–1174.
- Forestier E, Schmiegelow K. The incidence peaks of the childhood acute leukemias reflect specific cytogenetic aberrations. *J Pediatr Hematol Oncol* 2006; **28**: 486–495.
- Davies SM, Borowitz MJ, Rosner GL, Ritz K, Devidas M, Winick N *et al*. Pharmacogenetics of minimal residual disease response in children with B-precursor acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood* 2008; **111**: 2984–2990.
- Rocha JC, Cheng C, Liu W, Kishi S, Das S, Cook EH *et al*. Pharmacogenetics of outcome in children with acute lymphoblastic leukemia. *Blood* 2005; **105**: 4752–4758.
- Gregers J, Christensen IJ, Dalhoff K, Lausen B, Schroeder H, Rosthoj S *et al*. The association of reduced folate carrier 80G>A polymorphism to outcome in childhood acute lymphoblastic leukemia interacts with chromosome 21 copy number. *Blood* 2010; **115**: 4671–4677.
- Yang JJ, Cheng C, Devidas M, Cao X, Campana D, Yang W *et al*. Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. *Blood* 2012; **120**: 4197–4204.
- Wesolowska A, Dalgaard MD, Borst L, Gautier L, Bak M, Weinhold N *et al*. Cost-effective multiplexing before capture allows screening of 25 000 clinically relevant SNPs in childhood acute lymphoblastic leukemia. *Leukemia* 2011; **25**: 1001–1006.
- Schmiegelow K, Forestier E, Hellebostad M, Heyman M, Kristinsson J, Soderhall S *et al*. Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. *Leukemia* 2010; **24**: 345–354.
- Flohr T, Schrauder A, Cazzaniga G, Panzer-Grumayer R, van dV V, Fischer S *et al*. Minimal residual disease-directed risk stratification using real-time quantitative PCR analysis of immunoglobulin and T-cell receptor gene rearrangements in the international multicenter trial AIEOP-BFM ALL 2000 for childhood acute lymphoblastic leukemia. *Leukemia* 2008; **22**: 771–782.
- Conter V, Bartram CR, Valsecchi MG, Schrauder A, Panzer-Grumayer R, Moricke A *et al*. Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study. *Blood* 2010; **115**: 3206–3214.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
- Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
- Besag J, Mondal D. Exact goodness-of-fit tests for Markov chains. *Biometrics* 2013; **69**: 488–496.
- Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res* 2014; **42**: D472–D477.
- Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975; **405**: 442–451.
- Thorn CF, Klein TE, Altman RB. PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol Biol* 2013; **1015**: 311–320.
- Gottesman MM, Fojo T, Bates SE. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat Rev Cancer* 2002; **2**: 48–58.
- Gialeli C, Theocharis AD, Karamanos NK. Roles of matrix metalloproteinases in cancer progression and their pharmacological targeting. *FEBS J* 2011; **278**: 16–27.
- Sato Y, Goto Y, Narita N, Hoon DS. Cancer cells expressing Toll-like receptors and the tumor microenvironment. *Cancer Microenviron* 2009; **2**(Suppl 1): 205–214.
- Correa RG, Milutinovic S, Reed JC. Roles of NOD1 (NLRC1) and NOD2 (NLRC2) in innate immunity and inflammatory diseases. *Biosci Rep* 2012; **32**: 597–608.
- Toft N, Birgens H, Abrahamsson J, Bernell P, Griskevicius L, Hallbook H *et al*. Risk group assignment differs for children and adults 1-45yr with acute lymphoblastic leukemia treated by the NOPHO ALL-2008 protocol. *Eur J Haematol* 2013; **90**: 404–412.
- Frandsen TL, Heyman M, Abrahamsson J, Vetterranta K, Aasberg A, Vaitkeviciene G *et al*. Complying with the European Clinical Trials Directive while surviving the administrative pressure—an alternative approach to toxicity registration in a cancer trial. *Eur J Cancer* 2014; **50**: 251–259.
- Lund B, Wesolowska-Andersen A, Lausen B, Borst L, Rasmussen KK, Muller K *et al*. Host genome variations and risk of infections during induction treatment for childhood acute lymphoblastic leukaemia. *Eur J Haematol* 2014; **92**: 321–330.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>

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