HEMATOLOGY: RESEARCH ARTICLE

Pediatric Blood & Cancer

SIOP | aspho The American Society of Pediatric Hematology/Oncology

WILEY

Characteristics of white blood cell count in acute lymphoblastic leukemia: A COST LEGEND phenotype–genotype study

Marianne Helenius[1,2] | Goda Vaitkeviciene[3] | Jonas Abrahamsson[4] |
Ólafur Gisli Jonsson[5] | Bendik Lund[6] | Arja Harila-Saari[7] | Kim Vettenranta[8] |
Sirje Mikkel[9] | Martin Stanulla[10] | Elixabet Lopez-Lopez[11,12] | Esmé Waanders[13,14] |
Hans O. Madsen[15] | Hanne Vibeke Marquart[15] | Signe Modvig[15] |
Ramneek Gupta[1,16] | Kjeld Schmiegelow[2,17] | Rikke Linnemann Nielsen[1,2,16]

[1]Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Copenhagen, Denmark

[2]Department of Pediatrics and Adolescent Medicine, University Hospital Rigshospitalet, Copenhagen, Denmark

[3]Vilnius University Hospital Santaros Klinikos Center for Pediatric Oncology and Hematology and Vilnius University, Vilnius, Lithuania

[4]Department of Paediatrics, Institution for Clinical Sciences, Sahlgrenska University Hospital, Gothenburg, Sweden

[5]Department of Pediatrics, Landspitali University Hospital, Reykjavík, Iceland

[6]Department of Pediatrics, St. Olavs Hospital, Trondheim, Norway

[7]Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

[8]University of Helsinki and Childrent's Hospital, University of Helsinki, Helsinki, Finland

[9]Department of Hematology and Oncology, University of Tartu, Tartu, Estonia

[10]Department of Pediatric Hematology and Oncology, Hannover Medical School, Hannover, Germany

[11]Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), Leioa, Spain

[12]Pediatric Oncology Group, BioCruces Bizkaia Health Research Institute, Barakaldo, Spain

[13]Department of Genetics, University Medical Center Utrecht, Utrecht, The Netherlands

[14]Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands

[15]Department of Clinical Immunology, University Hospital Rigshospitalet, Copenhagen, Denmark

[16]Novo Nordisk Research Centre Oxford, Oxford, UK

[17]Institute of Clinical Medicine, Faculty of Medicine, University of Copenhagen, Copenhagen, Denmark

Correspondence
Rikke Linnemann Nielsen, Department of Health Technology, Technical University of Denmark, DK-2800 Kongens Lyngby, Copenhagen, Denmark.
Email: rlni@dtu.dk
Kjeld Schmiegelow, Department of Pediatrics and Adolescent Medicine, University Hospital Rigshospitalet, Copenhagen, Denmark.
Email: kjeld.schmiegelow@regionh.dk

Abstract

Background: White blood cell count (WBC) as a measure of extramedullary leukemic cell survival is a well-known prognostic factor in acute lymphoblastic leukemia (ALL), but its biology, including impact of host genome variants, is poorly understood.

Methods: We included patients treated with the Nordic Society of Paediatric Haematology and Oncology (NOPHO) ALL-2008 protocol (N = 2347, 72% were genotyped

by Illumina Omni2.5exome-8-Bead chip) aged 1–45 years, diagnosed with B-cell precursor (BCP-) or T-cell ALL (T-ALL) to investigate the variation in WBC. Spline functions of WBC were fitted correcting for association with age across ALL subgroups of immunophenotypes and karyotypes. The residuals between spline WBC and actual WBC were used to identify WBC-associated germline genetic variants in a genome-wide association study (GWAS) while adjusting for age and ALL subtype associations.

Results: We observed an overall inverse correlation between age and WBC, which was stronger for the selected patient subgroups of immunophenotype and karyotypes ($\rho_{BCP-ALL} = -.17, \rho_{T-ALL} = -.19; p < 3 \times 10^{-4}$). Spline functions fitted to age, immunophenotype, and karyotype explained WBC variation better than age alone ($\rho = .43, p << 2 \times 10^{-6}$). However, when the spline-adjusted WBC residuals were used as phenotype, no GWAS significant associations were found. Based on available annotation, the top 50 genetic variants suggested effects on signal transduction, translation initiation, cell development, and proliferation.

Conclusion: These results indicate that host genome variants do not strongly influence WBC across ALL subsets, and future studies of why some patients are more prone to hyperleukocytosis should be performed within specific ALL subsets that apply more complex analyses to capture potential germline variant interactions and impact on WBC.

KEYWORDS
acute lymphoblastic leukemia (ALL), genome-wide association studies (GWAS), genotype, spline functions, white blood cell count (WBC)

1 | INTRODUCTION

Acute lymphoblastic leukemia (ALL) is the most common childhood cancer and despite improvements of survival rates, it remains a major cause of death among children with cancer.[1] One of the historically strongest risk factors for treatment failure is a high initial white blood cell count (WBC) in peripheral blood at diagnosis, which for patients with WBC above the normal range has been used as a measure of tumor burden and extramedullary cell survival.[2–4] The widely used NCI criteria define a WBC cutoff at 50×10^9/L for risk grouping, but as it is challenged by the lack of a clear distribution mode/antimode at this dichotomous discriminator, many groups have replaced or complemented it with cytogenetics and minimal residual disease (MRD) during the first months of therapy.[2,5,6] A higher WBC is seen in children compared to adults, and an almost 10-fold higher median WBC is seen in patients with T-cell ALL (T-ALL) compared to B-cell precursor ALL (BCP-ALL). However, patients can have a WBC within the normal range and still have a high percentage of leukemic blasts in the bone marrow.[3,5] A high WBC reflects the leukemic cells' ability to survive outside the thymus and bone marrow, but it is unknown to which extent this reflects host genomic or acquired features of the leukemic cells. Patients with favorable cytogenetics like high hyperdiploidy (>50 chromosomes) or translocation t(12;21) typically have lower WBC, whereas the poorer prognosis *MLL*-rearranged patients will present with higher WBC.[3,7,8] Thus, using a crude WBC cutoff for risk stratification rather than one adjusted by host features of age, immunophenotype, and cytogenetics that have been previously associated with WBC levels may both under- and overestimate the potential impact of the tumor burden.[3,7,8] The benefit of more differentiated risk classifier models has been published for relapse prediction in childhood ALL but could potentially be improved by a better classification of WBC.[6]

Genome-wide association studies (GWAS) are applied to identify and quantify the strength of common single-nucleotide polymorphisms (SNPs) associating with the phenotype.[9–11] Normal WBC has previously been studied with GWAS, where people with abnormal WBC levels are excluded, showing germline variants influencing WBC and leukocyte subsets.[12–14]

We modeled the interactions of WBC risk factors across key childhood ALL subgroups to identify the underlying host genome variants influencing WBC at diagnosis of ALL.

2 | METHODS

2.1 | Study population and clinical data

The study cohort includes patients diagnosed with BCP-ALL or T-ALL. The Philadelphia chromosome-negative patients, aged 1–45 years, were diagnosed between July 2008 and February 2019. We included

patients without Down syndrome or other known leukemia predisposing syndromes. Patients were treated with the Nordic Society of Paediatric Haematology and Oncology (NOPHO) ALL-2008 protocol in the Nordic and Baltic countries (Denmark, Sweden, Norway, Finland, Iceland, Estonia, and Lithuania). A written consent was obtained for inclusion in GWAS. In the NOPHO ALL-2008 protocol, WBC at the time of diagnosis was used for initial treatment stratification for the induction therapy as a dichotomized parameter, where patients with BCP-ALL and WBC $<100 \times 10^9$/L were assigned to the non-high-risk group, whereas patients with T-ALL and/or WBC $\geq100 \times 10^9$/L were assigned to the high-risk group.[15] We used a natural logarithm to transform the WBC (logWBC) to obtain normally distributed data for modeling.

The NOPHO registry contains data of anthropometrics, diagnosis, and treatment on patients treated with the NOPHO ALL-2008 protocol.[15] The clinical information included age at diagnosis, sex, country, height, weight, karyotype, immunophenotype, DNA index, and WBC at diagnosis, as well as treatment-related information on MRD (see Supporting Methods) and risk stratification (standard [SR], intermediate [IR], or high risk [HR]) at treatment day 15 and end of induction therapy (EOI; day 29).

Clinical variables (Figure S2) with a maximum of 15% missing values were accepted before imputation using factor analysis for mixed data with the "missMDA" package in R v3.6.1.[16,17] MRD and information on risk stratification were not imputed, as this was used to correlate with WBC as measures of treatment outcome.

2.2 | Genotype data

Postremission DNA was collected from a total of 2050 patients, and written add-on consent was obtained for participation in genetic studies. The samples were genotyped in four batches using three versions of the Illumina Infinium Omni2.5exome-8-BeadChip arrays with 2,546,527–2,617,655 SNPs available per version.[18–20] A detailed description of the genotype data preprocessing is included in Supporting Information in Supplementary Methods.

2.3 | Spline functions to model WBC variation across ages and ALL subtypes

Hypothesizing that specific host genome variant will have the same proportional impact on WBC across key ALL subsets, third-degree spline functions were constructed to approximate a model of WBC distribution, given the variation of WBC by age, immunophenotype, and selected cytogenetics. Spline functions are piece-wise polynomials consisting of multiple polynomials separated by a set of limits on the axis of the independent variable, known as "knots," which create smaller windows for each polynomium.[21,22] We applied third-degree splines, meaning that third-degree polynomials are applied to fit the dependent variable logWBC between a set of knots to approximate complex distributions via a smooth curve. Spline functions were fitted to model logWBC as a function of age with three knots placed at 4, 7,

and 25 years of age, where the first knot corresponded to a previously known incidence peak of ALL and the second was placed to be near the end of the decrease in the incidence peak.[7,8,23] We fitted spline functions to the patients with BCP- and T-ALL, separately, as these have different distributions of WBC. Furthermore, several cytogenetic subsets of patients diagnosed with BCP-ALL have different levels of WBC,[3] where we adjusted the already fitted spline function by the difference in median logWBC of all the patients with BCP-ALL and median logWBC in the selected subgroups. The selected subgroups were BCP-ALL with t(1;19) (*E2A/PBX1* gene fusion) and t(12;21) (*ETV6/RUNX1* gene fusion) as well high hyperdiploidy (>50 chromosomes), *MLL* rearrangements, and the remaining BCP-ALL patients with none of the above (noted as "BCP-ALL other"), which were pooled as number of patients $N < 50$.[24] The spline residuals illustrate the individual variation of each patient separately from the variation attributed to known influences on WBC in ALL by its associations with age, immunophenotype, and cytogenetics. The spline residuals are defined as

$$\Delta = \log \text{WBC}_{spline} - \log \text{WBC}_{true}$$

The spline functions were fitted using Python v3.6.10 and the LSQUnivariateSpline function from the Python package scipy. interpolate v1.5.4.

2.4 | Statistical analysis and genome-wide association studies

Statistical testing of distributions and correlations were made in R v3.6.1[16] and Python v3.6.10 with the package scipy v1.5.4. The influence of data coming from different centers/countries and logWBC across the selected subgroups was inspected using ANOVA test. Correlations were tested with Spearman rank correlation coefficient (ρ), where $p < .05$ was considered significant. To investigate the patterns of WBC in children and adults, we compared the correlations of age versus logWBC and logWBC versus splines in groups of children and adults by thresholds at ages 15, 18, and 21. The thresholds are chosen to replicate from previous studies by Vaitkevičienė et al. (2013), Toft et al. (2013 and 2018), and Hu et al. (2020).[2,3,25,26]

PLINK2/1.90beta3[20] was used to perform GWAS with linear regression. Genotypes were assumed to have an additive effect on logWBC with respect to the number of minor alleles. A Bonferroni-corrected genome-wide significance of $p < 5 \times 10^{-8}$ was applied to investigate if significant associations could be detected, but we also manually assessed the results for interesting structures using Manhattan plots created with the qqman package in R v3.6.1.[16,27]

Top associated variants were annotated using the online version of Variant Effect Predictor (VEP)[28] from Ensembl GRCh37 release 103 with variants included within 10 kb upstream/downstream of gene boundaries. Gene annotations from VEP were further inspected with GeneCards[29] v5.1. Genotype-Tissue Expression (GTEx) Portal V8[30,31] was used to check if any of the top variants were significantly associated with expression in any tissue. RegulomeDB v2.0[32,33] was used to inspect intergenic/intronic variants for known functional impact.
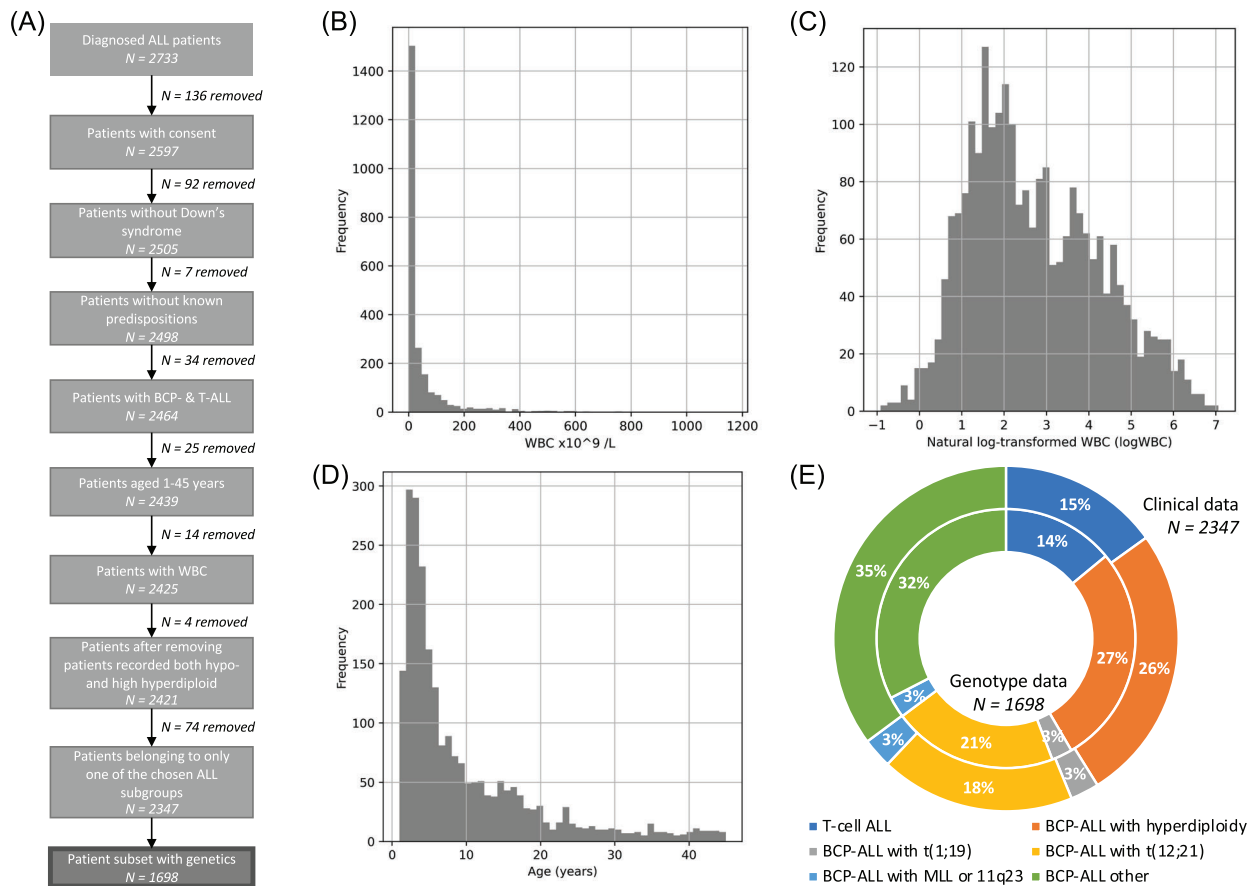
FIGURE 1 Overview of Nordic Society of Paediatric Haematology and Oncology (NOPHO) ALL-2008 cohort. (A) Flowchart describing the patient filtering for inclusion in this study. (B) Distribution of white blood cell count (WBC) measured from peripheral blood at acute lymphoblastic leukemia (ALL) diagnosis. (C) Distribution of WBC after natural logarithm transformation. (D) Distribution of age at ALL diagnosis. (E) Percentage distribution of selected ALL subgroups with clinical data ($N = 2347$) and genetic data ($N = 1698$)

We performed a gene set overlap analysis of the top gene annotations using the Molecular Signatures Databases (MSigDB) v7.4 and accompanying online tools.[34–36] The query gene set was constructed from the genes annotated to the top 50 SNPs in GWAS. We computed the overlap to the collections of gene sets from pathway databases "CP: KEGG"[37] and "CP: Reactome,"[38] as well as "C7: immunologic signature gene sets"[39] in MSigDB.

3 | RESULTS

3.1 | Characteristics of NOPHO ALL-2008 study cohort

The study included a total of 2347 patients after preprocessing and filtering (90% of diagnosed patients with consent; Figure 1A). MRD was measured at treatment days 15 and 29, where $N = 2235/2347$ patients (95.3% complete) were available for the study, though this differed for the individual days with $N_{day15} = 2139$ (91.1%) and $N_{day29} = 2164$ (92.2%) patients (Figure S3).

As the overall distribution was highly right skewed toward a lower WBC, we used a natural logarithm transformation to get more nor-

mally distributed data (Figure 1B,C). Age at diagnosis was right skewed toward a lower age with an incidence peak around 2–5 years (Figure 1D). The immunophenotypes were distributed as 85% patients with BCP-ALL and 15% with T-ALL (Figure 1E). We selected the following subtypes for further investigation: T-ALL and BCP-ALL with translocation t(1;19) or t(12;21), high hyperdiploidy, *MLL* rearrangement, as well as a subset termed "BCP-ALL other" (see groups in Figure 1E). The subsets we chose for correction with the spline functions were the largest in our cohort ($N = [63; 611]$; Table 1). "Extra cytogenetics" from Table 1 is a subset of "B-ALL other" and includes the smaller subgroups of ALL ($N = [7; 46]$) that were not included for the correction as separate groups, such as dicentric chromosome dic(9;20) or intrachromosomal amplification of chromosome 21 (iAMP21). The largest subgroups included were BCP-ALL with t(12;21) ($N = 428$), BCP-ALL hyperdiploidy ($N = 611$), and BCP-ALL other ($N = 825$), with the first two being the largest ALL subtypes.[40]

The median WBC was 9.0×10^9/L and 79.5×10^9/L for patients with BCP-ALL and T-ALL, respectively (Table 1). logWBC also differed significantly by karyotype ($p < 3 \times 10^{-4}$), with the main differences between those with an *MLL* rearrangement and the remaining BCP-ALL groups (Figure S4). A small inverse correlation between age and logWBC was observed across the full cohort ($\rho = -.06$, $p = 2.45 \times 10^{-3}$), which was

TABLE 1 Age and WBC at diagnosis of ALL for included patients

	Number of patients	Age (years)	WBC (×10⁹/L)	Spearman correlation between age and logWBC
All patients	2347 (72%)	5.72 (3.16 to 13.57, 1.01 to 44.94)	11.80 (4.6 to 46, .4 to 1161)	$\rho = -.06, p = 2.45 \times 10^{-3}$
Sex:				
Male	1345 (73%)	6.13 (3.27 to 14.76, 1.05 to 44.94)	12.30 (4.7 to 50.2, .4 to 1103)	$\rho = -.004, p = .89$
Female	1002 (71%)	5.27 (3 to 11.35, 1.01 to 44.94)	10.55 (4.4 to 42.6, .4 to 1161)	$\rho = -.15, p = 1.28 \times 10^{-6}$
Immunophenotype:				
BCP-ALL	1993 (73%)	5.09 (3 to 11.71, 1.01 to 44.94)	9 (4.2 to 31.6, .4 to 1161)	$\rho = -.17, p = 7.58 \times 10^{-15}$
T-ALL	354 (67%)	11.18 (6.14 to 20.03, 1.2 to 43.99)	79.45 (23.65 to 209.78, .6 to 1103)	$\rho = -.19, p = 2.46 \times 10^{-4}$
BCP-ALL subgroups:				
t(1;19) [E2A/PBX1]	63 (70%)	6.69 (3.41 to 13.08, 1.3 to 43.99)	18.4 (9.45 to 50.25, 1.4 to 306)	$\rho = -.44, p = 2.68 \times 10^{-4}$
t(12;21) [ETV6/RUNX1]	428 (82%)	3.94 (2.9 to 5.82, 1.22 to 20.08)	10.55 (5.28 to 32.65, .5 to 555)	$\rho = -.27, p = 1.62 \times 10^{-8}$
High hyperdiploidy	611 (77%)	4.12 (2.81 to 6.67, 1.1 to 44.6)	6.5 (3.5 to 17.9, .5 to 330)	$\rho = -.30, p = 3.52 \times 10^{-14}$
MLL rearrangement	66 (68%)	10.48 (2.01 to 25.64, 1.01 to 42.22)	99.5 (28.2 to 296.1, 2 to 1161)	$\rho = .22, p = .07$
BCP-ALL other	825 (67%)	9.33 (3.84 to 17.78, 1.05 to 44.94)	9.5 (4 to 36, .4 to 524)	$\rho = -.19, p = 5.57 \times 10^{-8}$
Extra cytogenetics:				
Hypodiploid (44 chr)	7 (71%)	2.81 (2.25 to 7.52, 1.65 to 25.87)	18.6 (7.25 to 128.65, 2 to 186)	$\rho = -.75, p = .05$
Hypodiploid (<44 chr)	27 (52%)	13.45 (5.69 to 20.43, 2.16 to 44.78)	6.7 (3.2 to 24.9, 1.1 to 119)	$\rho = -.16, p = .44$
dic(9;20)	38 (66%)	2.32 (1.74 to 4.79, 1.05 to 30.3)	45.9 (12.63 to 117.55, 1 to 374.1)	$\rho = -.27, p = .10$
iAMP21	46 (83%)	10.22 (7.53 to 14.39, 3.32 to 44.03)	8.05 (2.63 to 17.58, .6 to 218)	$\rho = -.19, p = .21$
CNS3 status (≥5 blast in cerebral spinal fluid)	93 (80%)	7.29 (3.16 to 13.05, 1.01 to 43.95)	44.1 (9 to 141, 1 to 1161)	$\rho = .07, p = .53$

Note: In parenthesis is the percentage patients for whom genetics was also available. The age and WBC columns are the median values per group with IQR and actual range (IQR, range). IQR is given as range first to third quantiles. The selected subsets of immunophenotype and cytogenetics used for correction in subsequent analysis are "T-ALL" and "Subgroups of BCP-ALL." Groups with significant correlation between age and WBC are in italic.
Abbreviations: ALL, acute lymphoblastic leukemia; BCP-ALL, B-cell precursor ALL; IQR, interquartile range; T-ALL, T-cell ALL; WBC, white blood cell count.

stronger for the immunophenotypes ($\rho_{BCP-ALL} = -.17$, $p_{BCP-ALL} = 7.58 \times 10^{-15}$; $\rho_{T-ALL} = -.19$, $p_{T-ALL} = 2.46 \times 10^{-4}$). For the subgroups investigated further, significant inverse correlations between age and WBC were found ranging $\rho = (-.44; -.19)$ (Table 1). Only the subgroup of BCP-ALL with *MLL* rearrangement had a nonsignificant positive correlation between age and WBC ($\rho = .22, p = .07$).

## 3.2 | Spline functions of logWBC across ages and ALL subtypes

We fitted two spline functions in this study: one for patients with T-ALL and one for BCP-ALL. These were fitted with preselected knots placed at ages of 4, 7, and 25 years, as seen to approximately fit with the distribution of age at diagnosis seen in our data (Figure 1D).

A trend of higher logWBC at a lower age, given logWBC as a function of age, is seen in Figure 2 and Table 2. For most subgroups, the splines approximately fit the trend of logWBC across ages, which is supported by a significant correlation between logWBC and the log-

WBC values estimated by the spline functions in the full cohort ($\rho = .43$, $p << 2 \times 10^{-16}$). Furthermore, the selected subsets of BCP-ALL with either high hyperdiploidy, translocations t(1;19) or t(12;21) showed distinct median logWBC levels, but similar logWBC–age trend, where we moved the fitted spline function (blue) to adjust for the variation between the ALL subgroups (orange). At the subgroup levels, the splines could, to some degree, capture the variation with significant correlations between logWBC and splines ranging $\rho = [.11; .46]$, where only the group of BCP-ALL with *MLL* rearrangement did not correlate. The correlations for ALL subgroups between splines and logWBC were similar compared to the subgroup correlations for age and logWBC, but at the cohort level allowed for a better capture of the complexity in the overall cohort ($\rho = -.06$ for age and logWBC in Table 1, vs. $\rho = .43$ for logWBC and splines in Table 2).

We investigated the correlations of age and WBC from previous studies and compared these to our study and spline adjustments to separately assess the patterns in children and adult cohorts. Vaitkevičienė et al. (2013)[3] studied $N = 2638$ patients with ALL aged 1–14.9 years and treated with the NOPHO ALL-92 and -2000 protocols, where
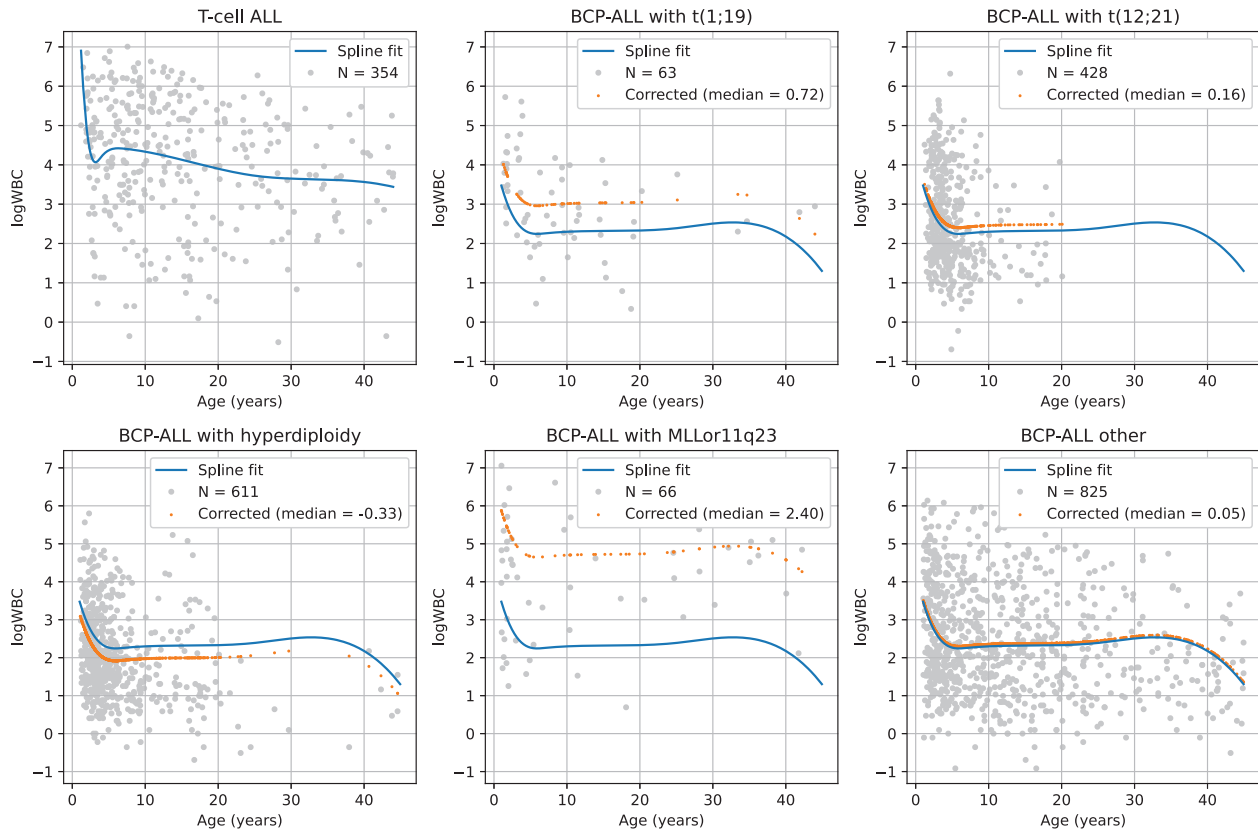
FIGURE 2 Spline functions with adjustment for subgroup median white blood cell count (WBC). The blue lines are the fitted spline functions. Gray dots are the patients' true values of age against natural log-transformed WBC (logWBC), while the orange dots are the true age against median-corrected spline values for logWBC

TABLE 2 Characteristics of spline functions and their correlations to logWBC

	Number of patients, n (%)	Spearman correlation between logWBC and splines	Spline residuals (IQR, range)
All patients	2347 (100%)	$\rho = .43, p = 1.12 \times 10^{-105}$	.19 (−.94 to 1, −3.89 to 4.75)
BCP-ALL	1993 (84.9%)	$\rho = .29, p = 1.84 \times 10^{-40}$	.25 (−.87 to 1, −3.89 to 4.09)
T-ALL	354 (15.1%)	$\rho = .23, p = 7.47 \times 10^{-6}$	−.18 (−1.16 to .93, −2.98 to 4.75)
BCP-ALL with t(1;19)	63 (2.7%)	$\rho = .46, p = 1.75 \times 10^{-}$	.28 (−.59 to .87, −2.49 to 2.7)
BCP-ALL with t(12;21)	428 (18.2%)	$\rho = .23, p = 1.95 \times 10^{-6}$	.31 (−.9 to 1, −3.88 to 3.12)
BCP-ALL with hyperdiploidy	611 (26%)	$\rho = .21, p = 2.51 \times 10^{-7}$	.31 (−.68 to .89, −3.32 to 2.69)
BCP-ALL with MLL rearrangement	66 (2.8%)	$\rho = −.02, p = .9$	.6 (−.86 to 1.69, −1.93 to 4.09)
BCP-ALL other	825 (35.2%)	$\rho = .11, p = 1.17 \times 10^{-3}$	.19 (−1.1 to 1.1, −3.89 to 3.55)

Note: Groups in italic are the selected subgroups that are used for correction with spline functions.
Abbreviations: ALL, acute lymphoblastic leukemia; BCP-ALL, B-cell precursor ALL; IQR, interquartile range; T-ALL, T-cell ALL; WBC, white blood cell count.

significant inverse correlations were found between age and WBC in some subsets of BCP-ALL ($\rho = [−.28; −.09]$, $p < .05$). We reported similar correlations in our full cohort aged 1–45 years (Table 1). The inverse correlation in patients with T-ALL ($N_{T\text{-ALL}} = 267, \rho_{T\text{-ALL}} = −.06$, $p_{T\text{-ALL}} = .33$) was not significant as opposed to this study ($\rho_{T\text{-ALL}} = −.19$, $p_{T\text{-ALL}} = 2.46 \times 10^{-4}$; Table 1). In our cohort, a significant correlation

between age and logWBC was also found in this range for children aged 1–15 years, whereas the corresponding population aged 15–45 years showed no correlation (Table 3).

In Toft et al. (2013),[2] a smaller set of patients ($N_{BCP\text{-ALL}} = 624$, $N_{T\text{-ALL}} = 125$) from the NOPHO ALL-2008 protocol is studied, and significant inverse correlations between age and WBC are reported

TABLE 3 Spearman correlations (ρ) of age-stratified groups from the NOPHO ALL-2008 study cohort ($N = 2347$)

Age group (years)	Number of patients	Spearman correlation of age and logWBC	Spearman correlation of logWBC and splines
1–15	1841 (78.44%)	$\rho = -.08, p = 1.14 \times 10^{-3}$	$\rho = .42, p = 2.49 \times 10^{-78}$
15–45	506 (21.56%)	$\rho = 0, p = .96$	$\rho = .45, p = 4.75 \times 10^{-27}$
1–18	1983 (84.49%)	$\rho = -.08, p = 4.30 \times 10^{-4}$	$\rho = .41, p = 1.18 \times 10^{-80}$
18–45	364 (15.51%)	$\rho = -.02, p = .77$	$\rho = .52, p = 4.11 \times 10^{-27}$
1–21	2069 (88.16%)	$\rho = -.08, p = 2.49 \times 10^{-4}$	$\rho = .41, p = 2.80 \times 10^{-83}$
21–45	278 (11.84%)	$\rho = -.10, p = .10$	$\rho = .58, p = 4.57 \times 10^{-26}$

Abbreviations: ALL, acute lymphoblastic leukemia; NOPHO, Nordic Society of Paediatric Haematology and Oncology; WBC, white blood cell count.

for the immunophenotypes ($\rho_{BCP-ALL} = -.2$, $\rho_{T-ALL} = -.3$; $p < .0001$). A second study by Toft et al. (2018)[25] reported with more power ($N_{BCP-ALL} = 1278$, $N_{T-ALL} = 231$) that the inverse correlations still hold in similar magnitude ($\rho_{BCP-ALL} = -.17$, $\rho_{T-ALL} = -.28$; $p < .001$). These are close to what we found in this study for the larger NOPHO ALL-2008 cohort ($N_{BCP-ALL} = 1993$ and $N_{T-ALL} = 354$, $\rho_{BCP-ALL} = -.17$, $\rho_{T-ALL} = -.19$; $p < 3 \times 10^{-4}$; Table 1). Considering only the children and young adults in the population aged 1–18 years, a significant inverse correlation is found, but not for the corresponding adult population aged 18–45 (Table 3).

As for the adult population, a previous study by Hu et al. (2020)[26] described significant associations between age and WBC in a healthy Chinese cohort ($N = 74,402$) aged 21–45 years; however, in our opinion, the association was limited, reflected by $\beta = [.01; .012]$ in a linear regression. For the age group 21–45 years in our (European) cohort, we have less power with $N = 278$ patients available and found a small insignificant inverse correlation between age and logWBC ($\rho = -.1$, $p = .1$; Table 3).

The corresponding age-stratified significant correlations between logWBC and splines were similar for children and adults, which emphasizes the fact that we have used age in the adjustment of WBC variation (Table 3).

The residuals between the actual and spline-predicted logWBC were estimated and used as a phenotype in GWAS. It was seen that most subgroups had positive residuals, except for patients with T-ALL (Table 2), which corresponds with the generally higher WBC levels seen in this subgroup. We associated the MRD status and values when positive (Figure S3) to the different representations of WBC in the form of logWBC and spline residuals (Table S1). The MRD status was significant for both WBC representations in the overall cohort at both day 15 and EOI. Yet, in the selected subgroups, only the MRD status at both days for BCP-ALL other and that at EOI for BCP-ALL with *MLL* rearrangement were significantly associated with both WBC representations. The significant associations were of similar effect size, though opposite directions between logWBC and spline residuals (odds ratio [OR] >1 and OR <1, respectively). Using the residuals enabled a representation of WBC variation accounting for some of the influences of age, immunophenotype, and selected karyotypes, whereby the GWAS focuses on only the unaccounted variance in its linear estimation of single genetic associations.

3.3 | Genome-wide association studies

A total of 2,146,366 variants and 1698 patients passed the genotype quality control and were included in our study cohort (72% of patients with clinical information; Figure 1 and Table 1). We conducted two GWAS using linear regression with the phenotypes logWBC and spline residuals of logWBC. The top 50 GWAS results from both phenotypes were annotated to closest genes and investigated with Gene Cards, GTEx, RegulomeDB, and gene overlap analysis.

3.3.1 | Genome-wide associations of logWBC

The first GWAS was conducted with logWBC as phenotype, because the linear regression model applied works under the assumption that the data are normally distributed. The GWAS did not show any genome-wide significant associations to logWBC. From the quantile–quantile (QQ) plot, the *p*-values observed were higher than expected with the deflation beginning before $-\log_{10}(p) = 3$ (genomic inflation factor $\lambda = .97$; Figure S5). Annotation of the top 50 associated variants found genes involved in signal transduction for cell development regulation, T helper cells, and neuronal differentiation (e.g., *CD81*, *CDH13*, *PTPRB*, *STAT4*, and *TM4SF5*; File S2).

3.3.2 | Genome-wide associations of logWBC spline residuals

The second GWAS was conducted with the before-mentioned spline residuals as the phenotype (Figure 2 and Table 2) and adjusted for the known influences of age, immunophenotype, and karyotype on logWBC (Figure S6). The most significant observed *p*-value was $p = 2.34 \times 10^{-6}$ for a rare SNP (rs144078525, G>A, minor allele frequency [MAF]$_{NOPHO} = 5.9 \times 10^{-3}$, MAF$_{EUR} = 2 \times 10^{-3}$; File S3) found in two heterozygous patients (one patient has T-ALL, the other has BCP-ALL with *MLL* rearrangement). There appeared to be a deflation in the *p*-values from the analysis, when inspecting the QQ plot of *p*-values from the GWAS, which was less deflation than seen in the previous GWAS (genomic inflation factor $\lambda = .99$; Figure S6). Annotation of the top 50

associated variants revealed genes involved in signal transduction and translation initiation, possibly in regulation of cell development, differentiation, and proliferation (e.g., *EIF6*, *ELK4*, *FAM83C*, *MYBBP1A*, and *TSPAN9*; File S3).

From GTEx, it was found that 12/50 SNPs were significantly associated with an altered expression in GTEx tissues. Of these, two SNPs on chromosome 20 were found significantly expressed in the whole blood (rs2425046 and rs6579227; File 3), which also presented with similar effect sizes in the GWAS ($\beta = -.37$ and $\beta = -.35$, respectively). With RegulomeDB, it was found that 17/50 SNPs had a probability larger than .5 of being a regulatory variant, whereof seven of 17 SNPs scored probabilities above .75 (rs2425046 and rs6579227 were also found here; File S3).

Our gene set overlap analysis with MSigDB from the top annotations found three significant gene set overlaps, of which one with five of 28 gene annotations was still significant after FDR correction at $\alpha = .05$ significance level ($p = 2.44 \times 10^{-7}$, FDR $q = 1.71 \times 10^{-3}$). The significant gene set was "GSE39820_CTRL_VS_TGFBETA3_IL6_IL23A_CD4_TCELL_DN"[41,42] from the collection "C7: immunologic signature gene sets"[39] including genes downregulated in CD4 T cells treated with TGF-β3, IL-6, and IL-32a. The significant overlap was the genes *C3*, *EDEM2*, *GPR108*, *HERPUD1*, and *TRIP10*, which annotated to three SNPs (rs201148371, rs2425046, and rs9938160; File S3) for which the genotypes distributed fairly equally across WBC in the ALL subgroups.

The two SNPs, rs6544982 and rs17146259, were found in the top 50 GWAS hits from both phenotypes (Files S2 and S3), which annotated to gene *BCYRN1* and intergenic region, respectively. The SNPs were found at similar ranking and *p*-values in the top of both GWAS, as well as similar effect sizes, though with opposite directions.

4 | DISCUSSION

The WBC pattern captured is age related as expected from associations to age found in previous studies of healthy cohorts and ALL patients, where also association between WBC with immunophenotype and cytogenetics is seen.[2,3,43–48] We assessed correlations by Spearman correlation coefficients (ρ) and considered the linear correlation of age and logWBC to be somewhat limited. This is both for pediatric (e.g., age 1–15 years: $\rho = -.08$, $p = 1.14 \times 10^{-3}$; Table 3) and for adults (age 15–45 years: $\rho = 0$, $p = .96$; Table 3). However, we see a much more significant correlation between logWBC and our fitted splines for both children (age 1–15: $\rho = .42$, $p = 2.49 \times 10^{-78}$) and adults (age 15–45: $\rho = .45$, $p = 4.75 \times 10^{-27}$). The correlations for both groups are very similar, highlighting the effectiveness of the spline function to capture the variation in age and cytogenetics (Tables 1 and 2).

As ALL is a very heterogeneous disease with many molecular subtypes and new ones are continuously being described, the importance of adjusting the clinical features by subtype is emphasized in this study,[7,49–51] which resembles what has recently been described for MRD.[6] Of note, some subsets, such as hypodiploidy and iAMP21, were not adjusted for due to the low frequency, and some of these are known to present with relatively high WBC (Table 1).[7] Furthermore, these were potentially present across our defined subgroups and thus influenced WBC in combination with these, or with other variations not accounted for in protocols.[40]

Our GWAS used a linear regression to fit and investigate the signals of SNPs on the target phenotype, measuring the variant effect on the phenotype with each addition of a minor allele. Using the spline residuals as phenotype in the GWAS makes the model try to account for variation of age, immunophenotype, and selected karyotypes. The spline functions are fitted with logWBC and the residuals thus interpreted on this scale as well. We see the residuals as still quite large across subgroups (Table 2), and from the QQ plot of the spline residuals GWAS (Figure S6), we see that little deflation of the *p*-values is still present. This may be due to the interpretation of the log-scaled phenotype applied to a linear model. However, the linear regression used in GWAS required that we used the log scale to satisfy the model assumption of normal distribution.

We found no genome-wide significant SNPs, but gene annotations of the top 50 associated SNPs suggested some effects on cell proliferation and differentiation, some of which also having been implicated in different cancers.[29,52–58] The top SNP rs144078525 ($\beta = 4.341 \pm .92$, $p = 2.34 \times 10^{-6}$, $MAF_{NOPHO} < .01$) reflects a missense variant in the second exon of *HOXC12*, which is part of the *HOXC* gene locus that contains a family of transcription factors expressed during embryo development and in lymphoid cell hematopoiesis,[59,60] and has been implicated in acute leukemias.[59,61] rs144078525 is also near the *HOXC13* gene.[28,29,57,59,62] Brotto et al. (2020)[57] associated *HOXC12* and *HOXC13* with cancer hallmarks for "Genome instability and DNA repair pathways" and "Tumor-promoting inflammation," respectively, in a pathway enrichment analysis. *HOXC12* expression has also been reported as significantly increased in breast tumor samples compared to normal by Luo et al. (2019).[62] *HOXC13* has been previously reported as fusion partner for NUP98 in patients with acute myeloid leukemia.[59,63] rs144078525 is also close to (5912 bp–7689 bp downstream) *HOTAIR*, which is a long noncoding RNA that regulates expression of the *HOXD* genes.[64] *HOTAIR* has been implicated in multiple cancers, including breast and colorectal, where it induces metastasis through the regulation of *HOXD*.[64]

rs6579227 ($\beta = -.30 \pm .07$, $p = 6.24 \times 10^{-6}$, $MAF_{NOPHO} = .13$) is 6418 bp–6880 bp upstream of *EIF6*, and rs2425046 ($\beta = -.28 \pm .07$, $p = 2.28 \times 10^{-5}$, $MAF_{NOPHO} = .13$) is an intronic variant in *EIF6*. *EIF6* encodes eukaryotic translation initiation factor 6, which is involved in the formation of the ribosomal 60S subunit and in translation initiation by prohibiting its joining with the 40S ribosome subunit to form the 80S active ribosome without the presence of mRNA.[55] *EIF6* has been reported overexpressed in multiple cancers, including colorectal, acute promyelocytic leukemia, and lung cancer metastasis.[58]

rs187168222 ($\beta = 1.63 \pm .39$, $p = 3.59 \times 10^{-5}$, $MAF_{NOPHO} < .01$) is a missense variant of *MYBBP1A*, which encodes MYB binding protein 1A that is involved in regulation of different transcription factors suggested to act as a tumor suppressor in a wide range of cellular process, including cell division, proliferation, and apoptosis.[56]

rs79050656 (β = .38 ± .09, p = 1.24 × 10^{-5}, MAF$_{NOPHO}$ = .07) and rs12370932 (β = .24 ± .06, p = 3.64 × 10^{-5}, MAF$_{NOPHO}$ = .18) are intronic variants of *TSPAN9*, which encodes cell surface protein tetraspanin 9.[29] Tetraspanin 9 belongs to a family of transmembrane proteins that mediates signal transduction in processes such as cell adhesion and invasion. Tetraspanin 9 has been reported in gastric cancer to inhibit the growth and invasion of tumor cells via the ERK1/2 signaling pathway.[29,65]

The variation of WBC at diagnosis of ALL is very diverse and shows indications of complex underlying feature interactions. Based on our models, it seems that the primary driver of WBC at diagnosis is the disease subtype of ALL, with single germline variants playing a very limited role. Still, the input data complexity or a combination thereof used in this study may be inadequate to explain the WBC variation. Consequently, a need remains for studies applying more complex models, on larger cohorts, and machine learning methods to provide opportunities to capture nonlinear biological interactions the potential is underlying WBC at diagnosis.[66,67] The study of genetic signals in biological complex phenotypes would benefit from utilizing the strengths offered by the machine learning methodologies both for discovery and prioritization of variants, as well as allowing for explorations of individual patient's risk factors determining their WBC.[68–70]

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Conception and design of the study: All authors. Analysis of the data: Marianne Helenius, Hans O. Madsen, Hanne Vibeke Marquart, and Signe Modvig. Interpretation of the data: Marianne Helenius, Goda Vaitkeviciene, Ramneek Gupta, Kjeld Schmiegelow, and Rikke Linnemann Nielsen. Drafted the manuscript: Marianne Helenius, Ramneek Gupta, Kjeld Schmiegelow, and Rikke Linnemann Nielsen. Revision and approval of final manuscript: All authors.

DATA AVAILABILITY STATEMENT

The data of this study are not available due to restrictions.

ORCID

Marianne Helenius https://orcid.org/0000-0003-3613-8338
Arja Harila-Saari https://orcid.org/0000-0003-2767-5828
Ramneek Gupta https://orcid.org/0000-0001-6841-6676
Kjeld Schmiegelow https://orcid.org/0000-0002-0829-4993
Rikke Linnemann Nielsen https://orcid.org/0000-0003-0173-2134

REFERENCES

1. Pui C-H, Yang JJ, Hunger SP, et al. Childhood acute lymphoblastic leukemia: progress through collaboration. *J Clin Oncol*. 2015;33:2938-2948.
2. Toft N, Birgens H, Abrahamsson J, et al. Risk group assignment differs for children and adults 1–45 yr with acute lymphoblastic leukemia treated by the NOPHO ALL-2008 protocol. *Eur J Haematol*. 2013;90:404-412.
3. Vaitkevičienė G, Forestier E, Hellebostad M, et al. High white blood cell count at diagnosis of childhood acute lymphoblastic leukaemia: biological background and prognostic impact. Results from the NOPHO ALL-92 and ALL-2000 studies. *Eur J Haematol*. 2011;86:38-46.
4. Smith M, Arthur D, Camitta B, et al. Uniform approach to risk classification and treatment assignment for children with acute lymphoblastic leukemia. *J Clin Oncol*. 1996;14:18-24.
5. Amin HM, Yang Y, Shen Y, et al. Having a higher blast percentage in circulation than bone marrow: clinical implications in myelodysplastic syndrome and acute lymphoid and myeloid leukemias. *Leukemia*. 2005;19:1567-1572.
6. Enshaei A, O'Connor D, Bartram J, et al. A validated novel continuous prognostic index to deliver stratified medicine in pediatric acute lymphoblastic leukemia. *Blood*. 2020;135:1438-1446.
7. Pain management. In: Pui C-H, ed. *Childhood Leukemias*. 3rd ed. Cambridge University Press; 2012. https://doi.org/10.1017/CBO9780511977633
8. Forestier E, Schmiegelow K, Nordic Society of Paediatric Haematology and Oncology NOPHO. The incidence peaks of the childhood acute leukemias reflect specific cytogenetic aberrations. *J Pediatr Hematol Oncol*. 2006;28:486-495.
9. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017;101:5-22.
10. Sud A, Kinnersley B, Houlston RS, Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer*. 2017;17:692-704.
11. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*. 2008;9:356-369.
12. Bao EL, Cheng AN, Sankaran VG. The genetics of human hematopoiesis and its disruption in disease. *EMBO Mol Med*. 2019;11:e10316.
13. Vuckovic D, Bao EL, Akbari P, et al. The polygenic and monogenic basis of blood traits and diseases. *Cell*. 2020;182:1214-1231.e11.
14. Crosslin DR, McDavid A, Weston N, et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Hum Genet*. 2012;131:639-652.
15. Schmiegelow K. Nordic Society of Paediatric Haematology and Oncology (NOPHO) ALL 2008 final protocol version 3a. Treatment protocol for children (1.0–17.9 years of age) and young adults (18–45 years

of age) with acute lymphoblastic leukemia. Nordic Society of Paediatric Haematology and Oncology; 2008. Accessed November 26, 2020. https://www.nopho.org

16. R Core Team. R: A Language and Environment for Statistical Computing. R Core Team; 2021.

17. Josse J, Husson F, missMDA: a package for handling missing values in multivariate data analysis. *J Stat Softw*. 2016;70:1-31. https://doi.org/10.18637/jss.v070.i01

18. Rayner W Strand home: genotyping chips strand and build files. Accessed April 22, 2020. https://www.well.ox.ac.uk/~wrayner/strand/index.html

19. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case-control association studies. *Nat Protoc*. 2010;5:1564-1573.

20. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.

21. Gauthier J, Wu QV, Gooley TA, Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. *Bone Marrow Transplant*. 2020;55:675-680.

22. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Springer US; 2021. https://doi.org/10.1007/978-1-0716-1418-1

23. Hjalgrim LL, Rostgaard K, Schmiegelow K, et al. Age- and sex-specific incidence of childhood leukemia by immunophenotype in the Nordic countries. *J Natl Cancer Inst*. 2003;95:1539-1544.

24. Iacobucci I, Mullighan CG, Genetic basis of acute lymphoblastic leukemia. *J Clin Oncol*. 2017;35:975-983.

25. Toft N, Birgens H, Abrahamsson J, et al. Results of NOPHO ALL2008 treatment for patients aged 1–45 years with acute lymphoblastic leukemia. *Leukemia*. 2018;32:606-615.

26. Hu W, Zhang P, Su Q, et al. Peripheral leukocyte counts vary with lipid levels, age and sex in subjects from the healthy population. *Atherosclerosis*. 2020;308:15-21.

27. Turner S. qqman: Q–Q and Manhattan plots for GWAS data. 2017.

28. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;17:122.

29. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics*. 2016;54:1.30.1-1.30.33.

30. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45:580-585.

31. Aguet F, Anand S, Ardlie KG, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369:1318-1330.

32. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22:1790-1797.

33. Dong S, Boyle AP. Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum Mutat*. 2019;40:1292-1298.

34. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Nat Acad Sci U S A*. 2005;102:15545-15550.

35. Liberzon A, Subramanian A, Pinchback R, et al. Molecular Signatures Database (MSigDB) 3.0. *Bioinformatics*. 2011;27:1739-1740.

36. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database hallmark gene set collection. *Cell Syst*. 2015;1:417-425.

37. Pathway Solutions. Accessed May 19, 2021. https://www.pathway.jp/

38. Reactome Pathway Database. Accessed May 19, 2021. https://reactome.org/

39. Godec J, Tan Y, Liberzon A, et al. Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity*. 2016;44:194-206.

40. Pui CH, Nichols KE, Yang JJ. Somatic and germline genomics in paediatric acute lymphoblastic leukaemia. *Nat Rev Clin Oncol*. 2019;16:227-240.

41. GSE39820_CTRL_VS_TGFBETA3_IL6_IL23A_CD4_TCELL_DN. Broad Institute Inc. Accessed May 19, 2021. http://www.gsea-msigdb.org/gsea/msigdb/geneset_page.jsp?geneSetName=GSE39820_CTRL_VS_TGFBETA3_IL6_IL23A_CD4_TCELL_DN

42. Lee Y, Awasthi A, Yosef N, et al. Induction and molecular signature of pathogenic TH17 cells. *Nat Immunol*. 2012;13:991-999.

43. Donadieu J, Auclerc MF, Baruchel A, et al. Prognostic study of continuous variables (white blood cell count, peripheral blast cell count, haemoglobin level, platelet count and age) in childhood acute lymphoblastic leukaemia. Analysis of a population of 1545 children treated by FRALLE. *Br J Cancer*. 2000;83:1617-1622. https://doi.org/10.1054/bjoc.2000.1504

44. Chmielewski PP, Strzelec B. Elevated leukocyte count as a harbinger of systemic inflammation, disease progression, and poor prognosis: a review. *Folia Morphol (Warsz)*. 2018;77:171-178.

45. Hulstaert F, Hannet I, Deneys V, et al. Age-related changes in human blood lymphocyte subpopulations: II. Varying kinetics of percentage and absolute count measurements. *Clin Immunol Immunopathol*. 1994;70:152-158.

46. Valiathan R, Ashman M, Asthana D. Effects of ageing on the immune system: infants to elderly. *Scand J Immunol*. 2016;83:255-266.

47. Vrooman LM, Blonquist TM, Harris MH, et al. Refining risk classification in childhood b acute lymphoblastic leukemia: results of DFCI ALL consortium protocol 05-001. *Blood Adv*. 2018;2:1449-1458.

48. Schmiegelow K, Forestier E, Hellebostad M, et al. Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. *Leukemia*. 2010;24:345-354.

49. Li J, Dai Y, Wu L, et al. Emerging molecular subtypes and therapeutic targets in B-cell precursor acute lymphoblastic leukemia. *Front Med*. 2021;15:347-371. https://doi.org/10.1007/s11684-020-0821-6

50. Lilljebjörn H, Fioretos T. New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood*. 2017;130:1395-1401.

51. Storti F, Moorman AV. New and emerging prognostic and predictive genetic biomarkers in B-cell precursor acute lymphoblastic leukemia. *Haematologica*. 2016;101:407-416.

52. Cipriano R, Miskimen KLS, Bryson BL, et al. Conserved oncogenic behavior of the FAM83 family regulates MAPK signaling in human cancer. *Mol Cancer Res*. 2014;12:1156-1165.

53. Snijders AM, Lee SY, Hang B, Hao W, Bissell MJ, Mao JH. FAM83 family oncogenes are broadly involved in human cancers: an integrative multi-omics approach. *Mol Oncol*. 2017;11:167-179.

54. Maurice D, Costello P, Sargent M, Treisman R. ERK signaling controls innate-like CD8[+] T cell differentiation via the ELK4 (SAP-1) and ELK1 transcription factors. *J Immunol*. 2018;201:1681-1691.

55. Hao P, Yu J, Ward R, et al. Eukaryotic translation initiation factors as promising targets in cancer therapy. *Cell Commun Signal*. 2020;18:175.

56. Felipe-Abrio B, Carnero A. The tumor suppressor roles of MYBBP1A, a major contributor to metabolism plasticity and stemness. *Cancers (Basel)*. 2020;12:254.

57. Brotto DB, Diogenes Siena AD, de Barros II, et al. Contributions of HOX genes to cancer hallmarks: enrichment pathway analysis and review. *Tumor Biol*. 2020;42:1010428320918050. https://doi.org/10.1177/1010428320918050

58. Ali MU, Ur Rahman MS, Jia Z, Jiang C. Eukaryotic translation initiation factors and cancer. *Tumor Biol*. 2017;39:1010428317709805. https://doi.org/10.1177/1010428317709805

59. Alharbi RA, Pettengell R, Pandha HS, Morgan R. The role of HOX genes in normal hematopoiesis and acute leukemia. *Leukemia*. 2012;27:1000-1008.

60. Bhatlekar S, Fields JZ, Boman BM. Role of HOX genes in stem cell differentiation and cancer. *Stem Cells Int*. 2018;2018:3569493.

61. Lawrence HJ, Fishbach NA, Largman C. HOX genes: not just myeloid oncogenes any more. *Leukemia*. 2005;19:1328-1330.

62. Luo Z, Rhie SK, Farnham PJ. The enigmatic HOX genes: can we crack their code? *Cancers (Basel)*. 2019;11:323.

63. Gough SM, Slape CI, Aplan PD. NUP98 gene fusions and hematopoietic malignancies: common themes and new biologic insights. *Blood*. 2011;118:6247-6257.

64. Wu Y, Zhang L, Wang Y, et al. Long noncoding RNA HOTAIR involvement in cancer. *Tumor Biol*. 2014;35:9531-9538.

65. Deng Y, Cai S, Shen J, Peng H. Tetraspanins: novel molecular regulators of gastric cancer. *Front Oncol*. 2021;11:2389.

66. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2009.

67. Balding DJ, A tutorial on statistical methods for population association studies. *Nat Rev Genet*. 2006;7:781-791.

68. Nicholls HL, John CR, Watson DS, et al. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Front Genet*. 2020;11:350.

69. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. *Front Genet*. 2019;10:267.

70. Upstill-Goddard R, Eccles D, Fliege J, Collins A. Machine learning approaches for the discovery of gene-gene interactions in disease data. *Brief Bioinform*. 2013;14:251-260.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Helenius M, Vaitkeviciene G, Abrahamsson J, et al. Characteristics of white blood cell count in acute lymphoblastic leukemia: A COST LEGEND phenotype–genotype study. *Pediatr Blood Cancer*. 2022;69:e29582. https://doi.org/10.1002/pbc.29582