Genetic susceptibility and pharmacogenetic factors in childhood acute lymphoid leukemia

PhD thesis

Orsolya Lautner-Csorba

Semmelweis University
Molecular Medicine Doctoral School

Supervisor:
Dr. Csaba Szalai

Opponents:
Dr. Katalin Boér
Dr. Zsolt Rónai

Chairman of the Comprehensive Examination Committee:
Dr. Mária Sasvári

Members of the Comprehensive Examination Committee:
Dr. Tamás Mészáros
Dr. Judit Kralovánszky

Budapest

2013

1. INTRODUCTION

Acute lymphoblastic leukemia (ALL) is the most frequent haematopoetic malignancy in childhood worldwide, and also in Hungary with about 70–80 new cases registered a year (Hungarian Children Cancer Registry). In the last years several genome wide and candidate gene association studies have tried to explore the genetic background of the disease, and revealed a number of genes and genetic variations, which might influence the risk of the disease or the response to the therapy. But these results are often inconsistent and require further studies for confirmation.

It is well known that genetic alterations of genes in transcription, lymphoid differentiation, DNA synthesis or in DNA methylation could be a first step for haematopoetic (lymphoid) malignant cell production.

The objective of this study was to investigate, whether genetic polymorphisms in genes related to xenobiotics detoxification, transcription factors, hematopoesis, cell proliferation and folate metabolic pathway influence the risk to childhood ALL. On the other hand we also studied, whether these variations influence the overall and event-free survival of the patients after the therapy. For this, we selected altogether more than hundred SNPs and nine SNPs more in *ABCC1* gene for pharmacogenetic study of anthracycline cardiotoxicity reagarding to childhood ALL.

Recently, several methods have been described and used for evaluating multiple interactions in multifactorial diseases (e.g. ALL). In the present study, for the statistical analysis we applied a traditional Frequentist method, and a Bayesian statistical methodology, named Bayesian network-based Bayesian multilevel analysis of relevance (BN-BMLA), which extends association analysis by estimating posteriors of strong relevance. In gene association studies it is well known, that the traditional Frequentist statistical methods have several limitations, like the difficult handling of the multiple testing problem and model complexity, the inappropriate approach towards complex traits, and the high redundancy of predictors (e.g. the discovery of non-causal, transitively associated descriptors). Bayesian networks in the Bayesian statistical framework offer an automated solution for this. By analyzing the data with BN-BMLA we intend to refine the dependency relationships of factors, e.g. we investigate whether a variable is directly relevant or its association is only mediated. BN-BMLA was recently extended with Bayesian effect size estimation providing a hybrid

measure, the Bayesian structure based odds ratio, which characterizes the parametric and strong relevance aspects of factors.

In this study we demonstrated that BN-BMLA can be a useful supplementary to the traditional Frequentist statistical methods in gene association studies.

2. OBJECTIVES

My aim was:

1. To examine the *GSTM1* and *GSTT1 null* variants, which are important in xenobiotic detoxification mechanism, and their role in the susceptibility of childhood acute lymphoid leukemia (ALL).

2. To analyse the relevance of the investigated 126 SNPs in 33 genes, that affect the process of hematopoesis, lymphoid differentiation, transcription, DNA synthesis…etc. between the patient and healthy control population. We also try to answer whether the related haplotypes of SNPs, and the different clinical parameters (e.g. gender, age, cytogenetics) affect the patogenesis of ALL in the whole, or in the B-cell, T-cell, hiperdiploid-ALL subgroups.

3. To reveal the role of the investigated gene variants in the same ALL population regarding to the overall (OS)-, and event-free (EFS) survival.

4. To determine the pharmacogenetic importance of *ABCC1* polymorphisms in the anthracyline-induced cardiotoxicity in childhood ALL.

5. The complementary application of our newly developed Bayesian network-based Bayesian multilevel analysis of relevance (BN-BMLA) method with the well-known Frequentist statistics during the data analysis. Based on our data we try to define the set of the relevant genetic and clinical factors for acute lymphoid leukemia and to estimate their *a posteriori* probability of relevance (P).

3. METHODS

Biobank

During the analyses we applied the biobank available in the Institute of Genetics, Cell-, and Immunobiology, Semmelweis University.

Patients were diagnosed with ALL between 1990 and 2010, aged approximately 1–15 years at diagnosis and treated according to the ALL Berlin-Frankfurt-Münster (BFM) 90, 95 and 2002 chemotherapy protocol. We stratified our patients in different risk groups according to the protocols' criteria. The control patients (approximately aged 16.1 ± 12.4 years) were randomly selected from healthy blood donors and from minor outpatients from the Orthopaedic Department in the Budai Children's Hospital, and from the Urological Department of Heim Pal Pediatric Hospital, Budapest. None of the controls have had childhood ALL or any other types of cancers previously. All study subjects belonged to the Hungarian (Caucasian) population. Informed consent was requested from the study subjects, or from the parents of patients. The study was conducted according to the principles expressed in the Declaration of Helsinki and was approved by the Hungarian Scientific and Research Ethics Committee of the Medical Research Council (ETT TUKEB; Case No.:8-374/2009-1018EKU 914/PI/08.)

DNA isolation

In a retrospective manner, DNA was obtained children who underwent chemotherapy due to acute lymphoblastic leukemia. The genomic DNA of ALL patients was obtained from peripheral blood in a retrospective manner using QIAmp DNA Blood Maxi Kit (Qiagen, Hilden, Germany). The healthy control DNA was isolated with iPrep PureLink gDNA Blood Kit, iPrep Purification Instrument (Invitrogen, Life Technologies Co., Grand Island, USA).

Selection of the investigated genes and polymorphisms

We selected gene variants from the results of GWA studies, candidate gene association studies and from other studies in which the investigated pathways could be important for ALL. In online databases (HapMap, OMIM, Genecards, dbSNP, UCSC Genome Browser, SNP Finder…etc.) we also searched for relevant SNPs. The selection criterion was: minor allele frequency > 10%. The SNPs were prioritized according to their published role in ALL, and their estimated functionality (1. missense, 2. promoter, 3. 3'-UTR, 4. coding-synonyme, 4.intron). In some cases (e.g. in *ARIDB5*, *IKZF1* genes) SNPs were selected, which showed strong linkage disequilibrium with other investigated SNPs in the HapMap database. This

could serve as genotyping controls, but during the BN-BMLA evaluation tag SNPs were chosen for the analysis.

The selected SNPs were classified into four main groups based on the aim of the research: candidate gene association study I, II, III, and pharmacogenetic study IV.:

I. group: gene variants which play role in the xenobiotic metabolism (*GSTM1, GSTT1*), 798 samples, 3 variants

II. group: gene polymorphisms which are affect the hematopoesis, lymphoid differentiation, DNA synthesis (e.g. *IKZF1, ARID5B, STAT3*..etc.), 1072 samples, 62 SNPs

III. group: gene polymorphisms in the folate pathway (e.g. *MTHFD1, MTRR, MTHFR*..etc.), 1072 samples, 64 SNPs

IV. group: single nucleotide polymorphisms of the *ABCC1* transporter molecule encoded gene, 235 samples, 9 SNPs

Genotyping of the investigated genes and polymorphisms

Single nucleotide polymorphisms (SNPs) were genotyped by Multiplex PCR reaction (*GSTM1, GSTT1*), Sequenom iPLEX Gold MassARRAY technology at the McGill University and Genome Quebec Innovation Centre, Montreal, Canada (SNPs of hematopoesis, DNA synthesis, folate metabolism…etc.), and with GenomeLab SNPstream system (*ABCC1*).

Examination of associations between clinical and genetic factors and the survival of acute lymphoid leukemia population

We also investigated which factors influence the 5- yearsurvival of the patients. We have data about the survival rate in 516 cases (95% of all patients). Our sample set contains similar rate of relapsed patients to what was observed in the whole population. Patients who died during the chemotherapy due to therapy resistant progressive disease or due to infections or toxicities of therapy are underrepresented in our sample. Survival analyses were performed using Kaplan-Meier method. The log-rank test was applied for evaluating the association between survival and categorized parameters, as risk group, gender, and study protocols. The statistical tests were carried out by IBM SPSS Statistics software, version 19.0.

Statistics

1. Frequentist method

Gender adjusted logistic regression model was applied to obtain odds ratios (ORs) and 95% confidence intervals (CIs) to estimate risks for each polymorphism to childhood ALL in the case-control data. Calculation of the genotype frequency and the allele positivity was performed by IBM SPSS Statistics software, version 21.0., and the allele frequency was analysed by Medcalc Version 12.4.0. The statistical analyses were performed not only for the

overall ALL population but also for the clinical subtypes (B-, T-cell and hyperdiplod (HD)-ALL). The Hardy-Weinberg equilibrium analysis was carried out by $\chi 2$ goodness-of-fit test through an online application (http://ihg. gsf.de/cgi-bin/hw/hwa1.pl). Multiple testing corrections were performed using the Benjamini- Hochberg false discovery rate (FDR) method with type I error rate of 5% ($p \leq 1{,}21E\text{-}03$) and of 1% ($p \leq 3{,}42E\text{-}04$). Thus, ''p'' values that could not reach the significance threshold of the false discovery rate of 5% or 1%, but above the $p \leq 0.05$ limit, were called nominally significant. The statistical power was calculated by R, using an own implementation of the Genetic Power Calculator of Purcell and Sham. The power calculations were adjusted using the same hypothesis correction method as described above. Linkage disequilibrium and haplotypes were calculated with Haploview version 4.2. Odds ratios for haplotypes were obtained by MedCalc 12.4.0 software. Survival analyses were performed using Kaplan-Meier method. The log-rank test was applied for evaluating the association between survival and categorized parameters, as risk group, gender, and study protocols. The statistical tests were carried out by IBM SPSS Statistics software, version 19.0.

2. Bayesian network-based Bayesian analysis of relevance (BM-BMLA)

The BN-BMLA method applies a Bayesian approach, which means that it computes the *a posteriori* probability of the strongly relevant variable sets with respect to a target variable (e.g. the variable describing the case/control status of the patients in a genetic association study). The strongly relevant variables have a direct influence on the target, thus these variables probabilistically shield the target from the effect of other variables. The *a posteriori* probability of the strong relevance (posterior) ranges from 0 to 1 and a posterior of 1 means that the target (e.g. immunophenotypes of ALL) definitely has a dependency relationship with a predictor (e.g. SNP), whereas 0 means there is no such relationship. Strong relevance has two types: direct relevance (e.g. a causal SNP) and pure interaction (e.g. a SNP with an epistatic effect). Besides strong relevance, other forms of structural associations exist. In this context, we examined statistical interaction and redundancy. In a structural approach, statistical interaction means that SNPs tend to occur together more often in the model as strongly relevant variables than it is expected according to the assumption of their independent relevance. On the other hand, redundancy means that two SNPs tend to be interchangeable in strongly relevant variable sets. The Bayesian structure based odds ratio is a recent extension to BN-BMLA which applies a hybrid approach towards effect size estimation by combining parametric relevance and strong relevance aspects. A Bayesian odds ratio is related to a specific target, and it is conditioned on corresponding dependency models

(i.e. graph structures representing the interactions of strongly relevant variables). The result is a posterior distribution over odds ratios, which provides a finer characterization of effect size and allows a more detailed analysis than a conventional confidence interval. Credible intervals (i.e. Bayesian analogue of confidence intervals) were computed based on the 95% HPD (high probability density) region of the Bayesian odds ratio.

4. RESULTS

1. Candidate gene association study I. – *GSTM1, GSTT1, CCR5*

Previous studies showed inconsistent results about the role of *GST* and *CCR5* variants in childhood leukemia. Based on our data, we found that the homozygous frequencies of *GSTM1*, and *GSTT1* null genotype in the control group and in patients with ALL, did not show significant difference (51.6% vs. 54.6%; p=0.399; OR=1.13 (95% CI (0.85-1.49)). Comparison of deletion allele frequencies of the *CCR5*Δ32 in control patient and ALL cases (8.8% vs. 8.8%; p=0.996; OR=1.00 (0.70-1.42)) showed neither significant difference. *GSTM1* or *CCR5* deletions or deletion both in *GSTT1* and *GSTM1*, or in *GSTM1* and *CCR5* genes, were not associated with ALL. We also examined the deletion combination of the adequate genes, the association between different genotypes, cytogenetics, immunphenotype and gender but found that these parameters did not affect the risk of childhood ALL.

2. Candidate gene association study II. – *ARID5B, IKZF1, STAT3*

Frequentist method: The differences in allele and genotype frequencies between ALL cases and controls were nominally significant for 20 SNPs. But, when gender adjusted logistic regression analysis with false discovery rate of FDR (α)=1% significance threshold was calculated, the differences remained significant (p<3.42E-04) only in cases of 6 SNPs in two genes (rs10821936, rs7089424 and rs4506592 in *ARID5B*, and rs6964969, rs11978267 and rs4132601 in *IKZF1*). These results indicate that these SNPs are associated with increased susceptibility to ALL with odds ratios between 1.4 and 1.5. Then we analyzed whether the number of risk alleles of each SNP influenced the susceptibility to ALL, and found that in all cases the homozygous states are associated with higher risk (OR between 1.9 - 2.1) than the carrier status, or heterozygous states. We calculated the linkage disequilibrium (LD) coefficients between the different SNPs, and found that in both genes, the significantly associated SNPs were in strong linkage with each other. This means that there is only one, but strong signal in each gene. We found two haplotypes, which influenced the susceptibility to ALL in the *IKZF1* gene, but the odds ratio associated with the haplotypes, were not higher than in the case of individual SNPs. When the hyperdiploid ALL was considered, two SNPs in the *STAT3* gene (rs3816769, rs12949918) showed decreased risk to this clinical subtype of the disease. We also investigated whether the gender of the patients influence the effect of the SNPs, but found no such effect.

BN-BMLA: Besides evaluating our results with the traditional Frequentist methods, we also analyzed them with our newly developed BN-BMLA method. Gender was also involved in the analysis as a variable. For each variable posterior probability for strong relevance was calculated. In case of ALL susceptibility, the most relevant SNPs are rs10821936 in *ARID5B* and rs4132601 in *IKZF1*. The probability that these SNPs are directly associated to ALL is 0.76 for rs10821936 and 0.97 for rs4132601, respectively. Both of these direct associations are even more probable in case of the B-cell lineage sample group (0.95 and 1.0 for rs10821936 and rs4132601, respectively). As the B-cell lineage is lot more frequent, the high probability of the strong relevance of these two SNPs in case of ALL susceptibility is probably due to their strong relevance in B-cell lineage. In the T-cell lineage, only the gender of the patient has a high probability of direct association to ALL susceptibility (0.85), where males have significantly higher odds of developing ALL than women (Frequentist OR = 2.28, C.I. 95%: 1.32 – 3.93). In the hyperdiploid (HD) sample group, the probabilities of the most relevant SNPs are moderate: rs12949918 in *STAT3* (0.60), rs12457893 in *BCL2* (0.57), rs3212713 in *JAK1* (0.56), and rs3087253 in *CCR5* (0.56). Detailed characterization of association relation showed multiple types of it. We computed the *a posteriori* probability of different association types with respect to ALL susceptibility in all sample groups. The probability of the association of the SNPs in *ARID5B* (rs4509706, rs4948487, rs10821936, rs4948496, rs4948502) to ALL susceptibility are equal or greater than 0.8, but only rs10821936 can be stated as directly relevant (0.76) and the posterior of strong relevance is below 0.13 in case of all other SNPs, which indicates their non-causal, non-functional role. The situation is same in case of the SNPs in the *IKZF1* gene. The association of the SNPs rs6954833, rs10235796, and rs4132601 in *IKZF1* to ALL susceptibility is highly probable (0.97), but only rs4132601 has a high probability (0.97) of being strongly relevant. These effects are more expressed in the B-cell lineage sample group. The probability of pure interaction is very low (below 0.1) in case of all SNPs in ALL susceptibility and B-cell lineage sample groups. However, in other sample groups, there are some SNPs with low probability of being in pure interaction to the phenotype, e.g. 0.43 in case of rs3212713 in *JAK3* in the hyperdiploid sample group, and 0.42 in case of rs2282883 in *AHR* when the target variable was the risk group of the patients. The Bayesian analysis offers a principled way to compute the measure of interaction or redundancy of two (or more) SNPs. We computed the redundancies and interactions between all variables in case of all sample groups. In case of ALL susceptibility, rs10821936 in *ARID5B* and rs17405722 in *STAT3* showed a weak interaction (Interaction ratio (IR) = 0.15) and two SNPs in *ARID5B*, namely

rs10821936 and rs4509706 showed a moderate redundancy (Redundancy ratio (RR) = 0.33). *ARID5B* and *IKZF1* (the two genes that have the highest posterior of strong relevance to ALL susceptibility) showed no interaction or redundancy with each other. In case of T-cell lineage sample group, the gender showed a weak interaction with three SNPs in three genes, namely rs703817 in *STAT6* (IR = 0.16), rs4987845 in *BCL2* (IR = 0.1), and rs1143684 in *NQO2* (IR = 0.11). This latter could be confirmed by logistic regression analysis, as well. This indicated, that male status increased the risk of T cell ALL. In case of the HD sample group, several overlapping components (i.e. sets of SNPs) were found that showed strong interaction. The component with the highest interaction ratio (IR = 3.06) consists of four SNPs in four genes, namely rs17405722 in *STAT3*, rs12457893 in *BCL2*, rs10208033 in *STAT1*, and rs3124603 in *NOTCH1*. The component with the second highest interaction ratio (IR = 2.72) consists of three SNPs in three genes, rs17405722 in *STAT3*, rs11888 in *JAK3*, and rs2030171 in *STAT1*. The third component (IR = 0.86) consists of three SNPs, rs3212713 in *JAK3*, rs3087253 in *CCR5*, and rs10235796 in *IKZF1*. Since these datasets are relatively small, exact characterization of these effects needs further validation.

3. Candidate gene association study III. – *Genes in the folate metabolism*

Frequentist method: The Frequentist statistical analysis revealed that 9 SNPs (rs2235013, rs12517451, rs1544105, rs1076991, rs12759827, rs9909104, rs2853533, rs3776455, rs1532268) in 8 genes (*ABCB1*, *DHFR*, *FPGS*, *MTHFD1*, *MTR*, *SHMT1*, *TYMS*, *MTRR*) reached the p,0.05 values, but due to the multiple testing, statistical corrections were applied. In this way 2 SNPs of the *MTHFD1* and *MTRR* genes reached the FDR =5% ($p \leq 1.21E-03$) significance threshold. The genotype distribution of the *MTHFD1* rs1076991 differed significantly between the overall ALL and control population (p = 1.94E-04; OR= 1.94 (1.37–2.76) for the GG genotype; power = 0.64). Subsequently, it was investigated whether this SNP was associated with the clinical characteristics of ALL: GG genotype increased the risk of B-cell ALL (p = 3.52E-04; OR= 2.00 (1.37–2.94); power= 0.59), but not of T-cell, or HD-ALL. Further evaluation of these results showed that the G allele positivity increased significantly the susceptibility to ALL both in allelic (50% vs. 41.8%; p= 1.30E-04; OR= 1.39 (1.18–1.65) and genotype levels (AA vs. AG/GG; p= 4.90E-04; OR= 1.61 (1.23–2.10)). Analyzing the ALL subgroups separately showed that this SNP increased the risk exclusively to the B-cell ALL. The frequency of the *MTRR* rs3776455 GG genotype differed between the ALL and control population (p= 4.49E-03; OR= 0.57 (0.38–0.84); power= 0.51). Further analysis of the genotype distribution showed that the GG genotype was associated with a significantly reduced risk to ALL (AA/AG vs. GG; p = 1.21E-03; OR= 0.55 (0.38–0.79)).

Haplotype analyses were carried out to study the influence of haplotype blocks of genes on ALL risk. Two haploblocks in the *MTHFD1* gene were found to influence the risk of ALL in our population. The ACTA haplotype had a slightly protective effect against ALL (p = 0.003; OR= 0.74 (0.61–0.90)). In contrast, people carrying the GCCA haploblock, had 1.38 fold greater risk (p =0.009; OR= 1.38 (1.08–1.76)) to ALL. Linkage disequilibrium coefficients (D', r2) were calculated for each of the significant SNPs in both genes (*MTHFD1* rs1076991, *MTRR* rs3776455) and found that the rs3776455 SNP in *MTRR* gene, was in strong linkage with other six SNPs (rs2966952, rs1801394, rs326120, rs1532268, rs162036, and rs10380) in the gene. It suggests that the rs3776455 SNP can determine the status of the other six SNPs, and thus can be regarded as a tagSNP.

BN-BMLA: We used the BNBMLA method to infer the posterior probability of strong relevance of the genetic markers with respect to ALL susceptibility. Similarly to the Frequentist analysis, the most relevant SNP in ALL susceptibility was the rs1076991 of the *MTHFD1* gene with (0.65). In B-cell lineage sample group, the probability of the strong relevance was 0.53, while in the T-cell lineage sample group, this probability was 0.13, and in the hyperdiploid sample group it was nearly zero. In the HD sample group, the relevant SNPs were rs3776455 and rs1532268 in the *MTRR* gene (probability of strong relevance 0.76 and 0.68), and rs1004474 in *TYMS* (0.66). According to the BN-BMLA the probability that rs2236225 and rs745686 in the *MTHFD1* gene were associated to ALL susceptibility was 0.71, but their probabilities of strong relevance to ALL were nearly zero while the value of their transitive relevance was 0.61, meaning that this association is probably mediated by the direct relevance of rs1076991 in the *MTHFD1* gene. The situation is the same in the case of *MTRR* gene. All measured SNPs (rs2966952, rs1801394, rs326120, rs1532268, rs162036, rs3776455 and rs10380) had around 0.65 probability of being associated to ALL, but only rs3776455 had a moderately high probability of strong relevance, meaning that the associations were mainly affected through this SNP. In the B-cell lineage sample group, three SNPs (rs1076991, rs2236225 and rs745686) in *MTHFD1* gene showed association to ALL with probabilities between 0.57–0.67, mediated by rs1076991. Both in the overall ALL and the B-cell lineage sample groups, all investigated SNPs in *SLCO1B1* (*SLC21A6*) were associated to ALL with probabilities between 0.51–0.55. In the case of the T-cell lineage sample group the SNPs in *TYMS* gene showed a relatively high probability of association (between 0.64–0.74), but only rs2853533 and rs1004474 SNPs had moderately high value of strong relevance, meaning that the associations were mediated through these SNPs. Similarly to the above detailed explanations, in the hyperdiploid sample group, the SNPs in the *TYMS*

gene were related to HD-ALL with probabilities of 0.72–0.77 mediated through the effect of rs1004474. SNPs in *MTRR* (rs2966952, rs1801394, rs326120, rs162036 and rs10380) were associated (0.79–0.82) via rs1532268 and rs3776455. We also examined the different types of interactions, namely structural interactions, statistical interactions and redundancies between the strongly relevant variables. In cases of the overall, B- and T-cell lineage sample groups, the analyses revealed no pure interactional effect. However, in the HD sample group some SNPs had moderately high probabilities for pure interaction (e.g. 0.68 in case of rs1532268 in *MTRR*). The SNPs rs1532268 in *MTRR*, rs1222809 in *DHFR*, rs11545078 and rs3780127 in *GGH* were in pure interaction with HD-ALL through rs1004474 in *TYMS*. Note, that directed edges represent only probabilistic relationships between the variables which are not necessary causal. In accordance with this, the interaction graph stated that for example rs1532268 in the *MTRR* gene was conditionally independent of HD-ALL, but the known genotype of rs1004474 in *TYMS* rendered it dependent. Therefore, the values of rs1004474 and its ‘‘parent’’ SNPs’ had a joint effect on the value of hyperdiploid ALL without a marginal effect of the ‘‘parents’’. We computed the statistical interactions and redundancies between all variables in case of all sample groups. In the whole patient group, rs1076991 in *MTHFD1* and rs3776455 in *MTRR* showed a weak redundancy (Redundancy ratio (RR) = 0.12). In the B-cell lineage sample group, rs1076991 in *MTHFD1* showed a weak redundancy with several SNPs in *SLCO1B1*, namely rs17328763 (RR =0.24), rs11045819 (RR = 0.16), rs11045818 (RR =0.13) and rs11045823 (RR = 0.11). In case of the T cell lineage sample group, the gender seemed to have a weak redundancy with rs10925257 in *MTR* (RR = 0.11). The rs1004474 in *TYMS*, rs1532268 in *MTRR*, rs1222809 in *DHFR*, and one of the two SNPs rs11545078 and rs3780127 in *GGH* (interaction ratio (IR) = 1.14, including rs11545078; and IR =1.12, including rs3780127) were in strong statistical interaction with each other. In line with this, the two SNPs in *GGH* showed a strong redundancy (RR = 0.85). It must be noted, however, that this dataset was relatively small, thus the exact characterization of these effects needs further validation. Bayesian odds ratios and corresponding credible intervals were investigated for relevant SNPs in order to characterize their effect on ALL. In case of *MTHFD1* rs1076991 the results confirmed its significant effect on the susceptibility to ALL. Both the AG and the GG genotype increased the risk of ALL with credible intervals of (1.44–1.52) and (1.88–2.01), respectively. This effect could also be observed in case of B-cell ALL patients. The credible intervals were similarly narrow (AG: 1.49–1.55, GG: 1.96–2.11). For T-cell ALL patients the *MTHFD1* rs1076991 had a similar effect although not as significant as in the previous cases. In contrast, the effect of *MTHFD1* rs1076991 was negligible in the

HD-ALL subgroup. The other SNP that was revealed to have a significant effect on the susceptibility to ALL was *MTRR* rs3776455. Based on the whole sample group the posterior distribution of its Bayesian odds ratio had multiple local maxima, resulting in disjoint credible intervals both for the AG (0.62–0.71, 0.96–1.13, 1.22–1.3) and the GG (0.34–0.6, 0.73–0.99, 1.25–1.39) genotypes. This means that possible dependency models could be divided into three groups each supporting different odds ratios. In case of the AG genotype the majority of the models (75%) supported the neutral odds ratio with a credible interval of (0.96– 1.13) which signified a non-relevant effect with relatively high probability. In case of the GG genotype however, the interval of (0.34–0.6) had the majority support of dependency models (84%), which confirmed the protective effect of the GG genotype with a high probability. On the other hand, the credible interval showing increased risk (1.25–1.39) also had some support (9%). This phenomenon might explain earlier results indicating the GG genotype as risk factor. Although the current study population indicates a protective effect, it is possible that the risk increasing effect could be observed under different circumstances. A similar structural uncertainty could be observed in the case of B-cell ALL patients, although the posterior distribution of Bayesian odds ratios was only bimodal. The effect of AG genotype of *MTRR* rs3776465 was neutral (0.99–1.04) with moderately high probability (0.75), whereas the GG genotype had a protective effect (0.53–0.54) with a similar probability (0.75). In contrast, in the HD-ALL subgroup the related Bayesian odds ratios of *MTRR* rs3776465 indicated strong effects with narrow credible intervals. The AG genotype increased the risk of HD-ALL (1.2–1.27), while the GG genotype had a remarkable protective effect (0.09–0.2). The interaction between *TYMS* rs1004474 and *MTRR* rs1532268 were also further investigated in this respect. On one hand, the *TYMS* rs1004474 had a negligible effect on the risk of HD-ALL both in case of GA and GG genotypes (with respect to AA) with credible intervals of (0.75– 0.97) and (0.94–1.10) respectively. On the other hand, both the GA and AA genotypes of *MTRR* rs1532268 led to a slightly increased risk to HD-ALL (1.28–1.40 and 0.97–1.72 with respect to GG), however the credible interval of the latter was exceedingly wide. Remarkably, the combination of rs1004474 (AA) - rs1532268 (GG) (i.e. the common homozygous case for both SNPs) had the highest conditional probability of risk of HD-ALL, namely 0.23, which exceeded the individual effect of risk allele A for *MTRR* rs1532268 with a conditional probability of risk of 0.14. Note, that whereas the univariate analysis indicated an increased risk for the carriers of the *MTRR* rs1532268 A allele, the multivariate analysis revealed a more complex pattern of interaction: the *MTRR* rs1532268 A allele became protective in carriers of the AA genotype of the *TYMS* rs1004474 with a joint OR of 0.32.

4. Pharmacogenetic study – *ABCC1*

Significant correlation was found between the reduction of left ventricular fractional shortening (LVFS; %) which is a marker of the anthracycline-induced cardiotoxicity, and the *ABCC1* rs3743527 TT (34%, p=0,001) variant compared to CC (39,5%) or CT (39,3%) genotype. Patients carrying the *ABCC1* rs3743527 mutant homozygous TT genotype had significantly reduced LVFS calculated previously from the echocardiography performed at the end of the anthracycline treatment phase of chemotherapy.

5. Examination of associations between clinical and genetic factors and the survival of acute lymphoid leukemia population

Frequentist method: The overall survival (OS) rate was 85.5% (n = 441) in our childhood ALL population. There was no significant difference between sex groups and study protocols in overall survivals. Significant difference was found between risk groups in overall survivals (p = 1E-7). The survival rates were 92.6% in the low, 87.0% in the medium and 62.3% in the high risk groups. The event-free survival (EFS) rate was 81.0%. There was also no significant difference between sex groups and study protocols in EFS. Risk groups in event-free survivals showed relevant differences (p=1E-07). The survival rates were 90.4% in the low, 82.6% in the medium and 60.4% in the high risk groups. Furthermore, we investigated, whether the SNPs in our study influenced the survival of the patients. Altogether 4 SNPs in 3 genes (rs10403561 and rs874966 in *CEBPA*, rs3024979 in *STAT6*, rs11667351 in *BAX*) were nominally significant in this respect, but none of them reached our significance threshold. Among these results the rs11667351 in the *BAX* gene showed the strongest association and gave the lowest p value (p=0.001). The OS and the EFS did not differ in this respect. An association between the OS rates and the rs9909104 SNP in the *SHMT1* gene of folate metabolism was also observed. OS of patients with CC (83.9%) was lower than the survival of the patients with major TT (88.8%) genotype (p=0.01).

BN-BMLA: In both cases, lineage (B- or T-cell) and risk group were found to be strongly relevant to the target variable (EFS, OS indicator) with high probability. However, it is more probable that lineage is in pure interaction (0.68 in EFS) than it has a direct relevance (0.09 in EFS). Its effect is mediated by the risk group, thus lineage is important only if risk group is known. In both cases, the strongly relevant genetic factors with the highest probability are rs11667351 in *BAX* (0.79 in EFS, and 0.87 in OS), and rs10403561 in *CEBPA* (0.63 in EFS, and 0.62 in OS). Besides, a SNP in *STAT6*, namely rs703817 can be indicated as strongly relevant in case of OS (0.67), but its probability is lower in case of EFS (0.34). We computed the redundancies and interactions between all variables in both survival types, as well.

5. CONCLUSIONS

1. We confirmed with both statistics (Frequentist and the Bayesian network-based Bayesian multilevel analysis of relevance=BN-BMLA), the role of genetic variations in *ARID5B* (rs10821936) and *IKZF1* (rs6964969) in the susceptibility to ALL, particularly to B-cell ALL in a relatively large Hungarian acute lymphoid leukemia population. The BN-BMLA analysis of the interactional relations showed that the *IKZF1* and *ARID5B* gene variants affect independetly the susceptibility of ALL, but the *ARID5B* and *STAT3* genes can act together.

2. With the BN-BMLA we proved that boys had significantly higher odds (2x) for developing ALL than girls.

3. Evaluation of the hyperdiploid (HD) subgroup with the Frequentist and BN-BMLA (Bayesian) methods, polymorphisms of *STAT3* rs12949918 proved to decrease the ALL risk. After the Bayesian analysis we also found a relevant set of polymorphisms (*BCL2* rs12457893, *JAK1* rs3212713, *JAK3* rs3212713, *CCR5* rs3087253) to hyperdiploid-ALL. It is also revealed that *AHR* (rs2282883) variant was in pure interaction with the risk group, which showed the indirect relevance of the SNP to HD-ALL.

4. In the case of T-cell lineage sample group, the gender showed interaction with the rs703817 in *STAT6*, rs4987845 in *BCL2*, and rs1143684 in *NQO2*. This indicated a gender-influenced mechanism in the development of ALL.

5. We found two haplotypes in the *IKZF1* gene, which influenced the susceptibility to ALL. TGGGG associated with increased (1.5x) and TGATA with (0.7x) decreased ALL risk.

6. In the folate metabolism the rs1076991 SNP in the *MTHFD1* gene and the rs3776455 in the *MTRR* gene significantly influenced the risk of ALL. Patients with *MTHFD1* rs1076991 GG increased the ALL risk, but carrying the *MTRR* rs3776455 GG genotype showed decreased susceptibility to ALL. We confirmed these findings with both statistics.

7. A particular feature of the gene-gene interactions was also revealed by the BN-BMLA. Carrying the *MTRR* rs1532268 SNP A allele were associated with a slightly increased risk to hyperdiploid ALL, while it turned to be protective when occurred together with the AA mutant genotype of the *TYMS* rs1004474 gene variant in the same metabolic pathway.

8. After the pharmacogenetic analysis we found that patients harboring the *ABCC1* rs3743527 TT genotype had significantly reduced left ventricular fractional shortening (LVFS) calculated at the end of the treatment compared to other genotype groups. This finding indicates the key role of *ABCC1* gene polymorphisms in the anthracyline-induced cardiotoxicity.

9. Evaluating the association of survival rates and the effect of the clinical and genetic factors, either the Frequentist and the BN-BMLA showed that risk group, lineage and genetic variations in the *BAX* and *CEBPA* genes might also influence the survival of the patients.

10. In the present study we demonstrated firstly on a relatively large Hungarian pediatric acute lymphoid leukemia population the several advantageous features of the BN-BMLA method, and we provided evidence that in gene association studies it might be a useful supplementary to the traditional Frequentist statistical method.

6. LIST OF SCIENTIFIC PUBLICATIONS

Publications directly related to the PhD thesis

Lautner-Csorba O, Gézsi A, Erdélyi DJ, Hullám G, Antal P, Semsei AF, Kutszegi N, Kovács GT, Szalai Cs. (2013) Roles of genetic polymorphisms in the folate pathway in childhood acute lymphoblastic leukemia evaluated by Bayesian relevance and effect size analysis. PLoS One, 8: e69843. IF: 3.730

Lautner-Csorba O, Gézsi A, Semsei Á, Antal P, Edélyi DJ, Schermann G, Kutszegi N, Csordás K, Hegyi M, Kovács G, Falus A, Szalai C. (2012) Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance. BMC Med Genomics, 5:42. IF: 3.466

Semsei AF, Erdelyi DJ, Ungvari I, Csagoly E, Hegyi MZ, Kiszel PS, **Lautner-Csorba O**, Szabolcs J, Masat P, Fekete G, Falus A, Szalai C, Kovacs GT. (2012) ABCC1 polymorphisms in anthracycline-induced cardiotoxicity in childhood acute lymphoblastic leukaemia. Cell Biol Int, 36: 79-86. IF: 1.640

Cumulative impact factor of publications directly related to the PhD thesis: 8.836

Other coauthored publications

Rausz E, Szilagyi A, Nedoszytko B, Lange M, Niedoszytko M, **Lautner-Csorba O**, Falus A, Aladzsity I, Kokai M, Valent P, Marschalko M, Hidvegi B, Szakonyi J, Csomor J, Varkonyi J. (2013) Comparative analysis of IL6 and IL6 receptor gene polymorphisms in mastocytosis. Br J Haematol, 160: 216-219. IF: 4.942

Hegyi M, Gulácsi A, Cságoly E, Csordás K, Eipel OT, Erdélyi DJ, Müller J, Nemes K, **Lautner-Csorba O**, Kovács GT. (2012) Clinical relations of methotrexate pharmacokinetics in the treatment for pediatric osteosarcoma. J Cancer Res Clin Oncol, 138: 1697-1702. IF: 2.914

Semsei Á, **Lautner-Csorba O**, Kutszegi N, Schermann G, Eipel O, Falus A, Szalai C, Kovács GT, Erdélyi DJ. (2012) A gyermekkori akut limfoid leukémia farmakogenetikája egy gyógyszermellékhatás példáján. Magy Tud, 173: 90-97.

Cumulative impact factor of all publications: 16.692