A Genome-wide Association Study of Susceptibility to Acute Lymphoblastic Leukemia in Adolescents and Young Adults

Virginia Perez-Andreu, MD PhD1*, Kathryn G. Roberts, PhD2*, Heng Xu, PhD1,3*, Colton Smith, PhD1, Hui Zhang, MD, PhD1, Wenjian Yang, PhD1, Richard C. Harvey, PhD4, Debbie Payne-Turner, BS2, Meenakshi Devidas, PhD5, I-Ming Cheng, DVM6, William L. Carroll, MD7, Nyla A. Heerema, PhD8, Andrew J. Carroll, PhD9, Elizabeth A. Raetz, MD10, Julie M. Gastier-Foster, PhD11, Guido Marcucci, MD12, Clara D. Bloomfield, MD12, Krzysztof Mrózek, MD, PhD12, Jessica Kohlschmidt, PhD12,13, Wendy Stock, MD14, Steven M. Kornblau, MD15, Marina Konopleva, MD, PhD16, Elisabeth Paietta, PhD17, Jacob M. Rowe, MD18, Selina M. Luger, MD19, Martin S. Tallman, MD20, Michael Dean, PhD21, Esteban G Burchard, MD22, Dara G. Torgerson PhD22, Feng Yue, PhD23, Yanli Wang23, Ching-Hon Pui, MD24, Sima Jeha, MD24, Mary V. Relling, PharmD1, William E. Evans, PharmD1, Daniela S. Gerhard, PhD25, Mignon L. Loh, MD26, Cheryl L. Willman, MD6, Stephen P. Hunger, MD27, Charles G Mullighan, MD2 and Jun J. Yang, PhD1

1Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN; 2Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN; 3State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, West China Medical School, Sichuan University, Chengdu, China; 4University of New Mexico Cancer Center, Albuquerque, NM; 5Department of Biostatistics, Colleges of Medicine, Public Health & Health Professions, University of Florida, Gainesville, FL; 6Cancer Research and Treatment Center, University of New Mexico, Albuquerque, NM; 7Laura and Isaac Perlmutter Cancer Center, New York University Langone Medical Center, New York, NY; 8Department of Pathology, College of Medicine, The Ohio State University, Columbus, OH; 9Department of Genetics, University of Alabama at Birmingham, Birmingham, AL; 10Huntsman Cancer Institute, The University of Utah, Salt Lake City, UT; 11The Ohio State University, Nationwide Children's Hospital, Columbus, OH; 12Comprehensive Cancer Center, The Ohio State University, Columbus, OH; 13Alliance of Clinical Trials in Oncology Statistics and Data Center, Mayo Clinic, Rochester; 14Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago, IL; 15Section of Molecular Hematology & Therapy, Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX; 16The University of Texas MD Anderson Cancer Center, Houston, TX; 17Department of Medicine (Oncology), Albert Einstein College of Medicine, Yeshiva University, NY, NY; 18Rambam Medical Center, Haifa, Israel; 19Hematologic Malignancies Program, Hematology-Oncology Division, University of Pennsylvania, Philadelphia, PA; 20Leukemia Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY; 21Laboratory of Experimental Immunology, National Cancer Institute, Frederick, Maryland; 22Department of Bioengineering & Therapeutic Science and Department of Medicine, University of California, San Francisco, San Francisco, California; 23Department of Biochemistry and Molecular Biology, Pennsylvania State University College of
RUNNING HEAD GWAS of susceptibility to ALL in AYA.

KEY POINTS

1) In this first ALL GWAS in AYAs, we identified that inherited GATA3 variants strongly influence ALL susceptibility in this age group.

2) These findings revealed similarities and differences in genetic basis of ALL susceptibility between young children and AYAs
ABSTRACT

Acute lymphoblastic leukemia (ALL) in adolescents and young adults (AYA: 16-39 years) is characterized by distinct presenting features and inferior prognosis compared to pediatric ALL. To better understand the disease etiology in this age group, we performed a genome-wide association study (GWAS) to comprehensively identify inherited genetic variants associated with susceptibility to AYA ALL. In the discovery GWAS, we compared genotype frequency at 635,297 SNPs in 308 AYA ALL cases and 6,661 non-ALL controls by using a logistic regression model with European, African and Native American genetic ancestries as covariates. SNPs that reached association \( P \leq 5 \times 10^{-8} \) in the discovery GWAS were tested in an independent cohort of 162 AYA ALL cases and 5,755 non-ALL controls. We identified a single susceptibility locus on 10p14 signified by two SNPs within the GATA3 gene with genome-wide significant associations: rs3824662, \( P = 2.8 \times 10^{-10} \), odds ratio (OR)=1.77 and rs3781093, \( P = 3.2 \times 10^{-9} \), OR=1.73. These findings were validated in an independent replication cohort. The risk allele at rs3824662 was most frequent in Philadelphia chromosome (Ph)-like ALL but also conferred susceptibility to non-Ph-like ALL in AYAs. In 1,827 non-selected ALL cases, the risk allele frequency at this SNP was positively correlated with age at diagnosis (\( P = 6.29 \times 10^{-11} \)). As the first GWAS of ALL susceptibility in the AYA population, our results point to unique biology underlying leukemogenesis and potentially distinct disease etiology by age group.
INTRODUCTION

Cancer survival rates have been steadily increased in the US across age groups except for adolescents and young adults (AYA), partly due to the persisting inferior treatment response in hematologic malignancies. Particularly with acute lymphoblastic leukemia (ALL), age as a continuous variable is negatively correlated with prognosis, in spite of risk-adapted combination chemotherapy. In an analysis of 21,626 ALL cases diagnosed between 1990 and 2005 and treated on Children’s Oncology Group (COG) frontline protocols, survival rates decreased significantly with increasing age at diagnosis regardless of treatment era (e.g., 94.1% for age 1-10, 84.7% for age 10-15, and 75.9% for age 15-22 in the 2000-2005 cohort). While pediatric-based treatment regimens have been tested in AYA populations and resulted in improved survival, the gap in treatment outcome between age groups persists and ALL remains one of the leading causes of cancer-related deaths in the AYA population.

The inferior prognosis of AYA is likely to be multifactorial, including socioeconomic factors, medication adherence, clinical trial enrollment, and importantly age-related differences in ALL tumor and host biology. For example, as age increases, there is a progressive rise in prevalence of ALL genetic subtypes with poor prognosis (e.g., Philadelphia chromosome-positive (Ph+)7, Ph-like8 or intrachromosomal amplification of chromosome 219), whereas subtypes with favorable outcome (high hyperdiploidy10 and ETV6-RUNX111) become less common. These comparisons are informative but also limited because they are primarily driven by ALL features discovered in children and/or older adults. As a result, differences in tumor biology between AYA and childhood ALL may have been underestimated and genomic profiling studies focusing on the AYAs are likely to reveal novel molecular features unique to this population.
Inherited genetic variations can strongly influence both the susceptibility to ALL\textsuperscript{12-15} and treatment outcomes\textsuperscript{16-20}. For example, genome-wide association studies (GWAS) have identified germline genetic variants at \textit{ARID5B}, \textit{IKZF1}, \textit{CEBPE}, \textit{PIP4K2A} and \textit{CDKN2A/CDKN2B} loci with substantial cumulative effects on ALL disease risk in children. These ALL susceptibility genes are involved in lymphoid cell development, cell cycle control, and tumor suppression, collectively affecting leukemogenesis. While one’s inherited genetic variants remain unchanged over life time, it is possible that the effects of these susceptibility variants vary by age, thus contributing to the age-related differences in ALL incidence and subtype. In fact, when we examined the risk of ALL conferred by the \textit{ARID5B} variant, there was a clear trend of diminishing effects with increasing age (i.e., allelic odds ratio of 2.01, 1.8, and 1.48 in children younger than 5, 5-10, and older than 10, respectively)\textsuperscript{15}. However, germline variants related to ALL risk in the AYA population have not been comprehensively examined.

To better understand the potential unique leukemia etiology in AYAs, we conducted the first GWAS to systemically interrogate germline single nucleotide polymorphisms (SNPs) for their contribution to ALL risk in this age group.
METHODS

Study design and patients

In the discovery GWAS, the ALL cases comprised 209 adolescents (median 17.4, range 16-21 years) and 99 young adults (median 24.3, range 21-39 years) with newly-diagnosed B- ALL, who were treated on the Children’s Oncology Group (COG; N=202)\textsuperscript{21}, the Alliance-Cancer and Leukemia Group B (N=56)\textsuperscript{22}, Eastern Cooperative Oncology Group E2993 (N=29)\textsuperscript{23}, MD Anderson Cancer Center\textsuperscript{24-27} (N=11), and St. Jude Children’s Research Hospital trials (N=10)\textsuperscript{28}. The subjects were chosen based on the availability of germline DNA which was extracted from peripheral blood samples during clinical remission (<5% blasts cells in bone marrow). A total of 6,661 unrelated subjects from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort (dbGaP phs000209.v9) were considered as non-ALL control subjects because the prevalence of adult survivors of childhood ALL is extremely low\textsuperscript{14}.

For the replication analyses, 162 children with ALL aged 16-21 from the COG P9900 protocols (COG P9905: NCT 00005596; COG P9904: NCT 00005585\textsuperscript{29}; COG P9906: NCT 00005603\textsuperscript{30} and COG AALL0232: NCT 00075725\textsuperscript{21}) were included. A set of 5,755 unrelated non-ALL controls were included in the replication analysis: 1,228 African Americans (AAs) from the AIDS Linked to Intravenous Experience cohort (ALIVE)\textsuperscript{31}, 880 Hispanic Americans from the Genetics of Asthma in Latino Americans study (GALA)\textsuperscript{32} and 3,647 European Americans from the Genetic Association Informative Network schizophrenia cohort (dbGAP phs000021.v3.p2)\textsuperscript{33} and GAIN bipolar cohort (phs000017.v3.p1\textsuperscript{34}, Figure 1).

The clinical trials were approved by local institutional review boards (IRB) and informed consent for trial enrollment and banking of specimens for future research were obtained from parents, guardians, or patients, as appropriate. This study was approved by the St. Jude Children’s Research Hospital IRB.
Genotyping and quality control

Genome-wide SNP genotyping was performed by using the Affymetrix Human SNP Array 6.0 for ALL cases in the discovery GWAS, those in the COG P9905, P9904 and AALL0232 cohorts, and for all non-ALL controls subjects (dbGaP MESA, ALIVE, GAIN and GALA). Genotype calls (coded as 0, 1, and 2 for AA, AB, and BB genotypes) were determined by the Birdseed v2 (Affymetrix SNP 6.0) algorithm\textsuperscript{35}. Samples for which genotypes were ascertained for less than 95\% of SNPs on the array were deemed to have failed and were excluded from the analyses. For the ALL cases in the COG P9906, genome-wide SNP genotyping was performed by using Affymetrix Human SNP Array 500K and GATA3 SNPs were genotyped by PCR and Sanger sequencing, as described previously\textsuperscript{36}. We did not observe evidence of potential genotyping errors in the germline DNA due to tumor cell contamination (data not shown).

Prior to GWAS, SNPs were subjected to a series of quality control steps (Supplemental Figure 1). First, we filtered SNPs on the basis of minor allele frequency (MAF) and SNP call rate: for SNPs with a MAF of 1\%-3\%, we excluded those with a call rate <99\%; for SNPs with MAF of 3\%-5\%, we excluded those with a call rate <98\%; for SNPs with a MAF >5\%, we excluded those with a call rate <95\%. An additional filtering step was applied in the GWAS involving non-ALL controls: we removed SNPs for which genotype frequencies differed significantly among control groups (i.e. dbGaP MESA vs. HapMap unrelated CEU or dbGaP MESA vs. the GAIN bipolar cohort [dbGaP phs000017.v3]\textsuperscript{33}, \(P<10^{-6}\) by \(\chi^2\) test and the comparison was restricted to European Americans (EAs). Finally, those SNPs deviating from Hardy-Weinberg equilibrium (\(P<0.01\), in EAs cases or controls) were also excluded from the analysis. After quality control filters were applied, 635,297 SNPs were included in the GWAS.
Genetic ancestry and population structure

Genetic ancestry was determined by using STRUCTURE (version 2.2.3)\textsuperscript{37}, on the basis of genotypes at 30,000 SNPs randomly selected from the Affymetrix SNP arrays. HapMap samples from descendants of Northern Europeans (CEU, N=60), West Africans (YRI, N=60), East Asians (CHB/JPT, N=90), and Native American references (NA; N=105)\textsuperscript{38} were used to represent European, African, Asian and NA ancestries, respectively. We assumed that these four ancestries summed to 100\% in each genotyped individual. EAs, AAs, NAs, and Asians were defined as having >95\% European genetic ancestry, >70\% African ancestry, >90\% NA ancestry, and >90\% Asian ancestry, respectively. Hispanics were individuals for whom NA ancestry was >10\% and greater than African ancestry (including genetically-defined NAs). The rest of the subjects were grouped as “Others”.

We also performed principal component analysis of ALL cases and controls in the discovery GWAS cohort including all SNPs that passed the quality control, and observed comparable population structure between cases and controls (Supplemental Figure 2). In addition, we exhaustively examined potential relatedness within ALL cases and within controls included in the discovery GWAS by computing pairwise identity by descent probabilities. No evidence of first or second-degree relationships was identified.

ALL somatic genomic lesions

In the discovery cohort the ALL genetic subtypes included high hyperdiploid (>50 chromosomes), \textit{ETV6-RUNX1}, \textit{TCF3-PBX1}, MLL-rearranged, \textit{BCR-ABL1} (Ph\textsuperscript{+}), and Ph-like (with or without \textit{CRLF2} rearrangements). Ph-like and \textit{ERG}-deregulated ALL were defined by Predictive Analysis of Microarrays\textsuperscript{21,39}. In the COG P9900 series, ALL subtypes included \textit{ETV6-RUNX1}, \textit{TCF3-PBX1}, hyperdiploid, and MLL-rearranged, with the remainder of cases

\[8\]
considered as B-other. GATA3 expression was quantified in 237 ALL blasts in 237 AYA cases, using Affymetrix U133A array.

**Statistical analysis**

In the discovery GWAS, the association test between genotypes at each of the 635,297 SNPs and ALL susceptibility was tested by comparing genotype frequency between AYA ALL cases and non-ALL controls using a logistic regression test under an additive model, including European, African and NA ancestry (as continuous variables) as covariates using PLINK (v1.07). Population stratification was assessed by the construction of a quantile-quantile plot (Supplemental Figure 3), and there was only a minimal inflation at the upper tail of the distribution (λ=1.02). SNPs that reached the association \( P \leq 5 \times 10^{-8} \) in the discovery GWAS were evaluated in the independent replication series (one-tailed test). In both discovery and replication groups, we also tested GATA3 SNPs separately in EAs, AAs, and Hispanic Americans.

R (version 2.15.1) statistical software was used for the rest of the analyses unless indicated otherwise. Statistical tests were chosen as appropriate and according to the phenotype distribution (e.g., normally or binomially distributed for continuous or categorical variables, respectively). Associations of SNP genotype with somatic lesions and age were estimated by logistic regression and linear regression test, respectively, after adjusting for genetic ancestry. Associations of GATA3 SNP genotype with GATA3 gene expression was assessed by linear regression model, adjusting for genetic ancestry.
RESULTS
AYA ALL GWAS

In the discovery GWAS, we compared genotype frequency at 635,297 SNPs between 308 AYA ALL cases and 6,661 non-ALL controls (Figure 1). After adjusting for genetic ancestry, only two SNPs at 10p14 within the GATA3 gene reached genome-wide significance: rs3824662 (P=2.84×10^{-10}, OR=1.77 [95%CI, 1.48 to 2.12]) and rs3781093 (P=3.20×10^{-9}, OR=1.73 [1.44 to 2.08], Table 1 and Figure 2). These two SNPs were in strong linkage disequilibrium (LD, r^2=0.94, D'=1 in HapMap CEU; Supplemental Figure 4), representing a single susceptibility locus. The A allele at rs3824662 was significantly over-represented in ALL cases compared to controls (35% vs. 20%), and consistent across race/ethnicity (i.e., EAs [30% vs. 17%, P=1.09×10^{-5}], Hispanics [50% vs. 33%, P=0.0008] and AAs [20% vs. 10%, P=0.07], Figure 3A). rs3781093 was significantly associated with ALL risk in EAs and Hispanics, but not in individuals of African descent in whom it was no longer in LD (r^2=0.006, D'=0.16) with rs3824662 (Supplemental Figure 5A).

To validate the association signals at these GATA3 SNPs, we tested in an independent set of 162 AYA ALL cases enrolled in COG P9900 and ALL0232 protocols and additional 5,755 non-ALL controls. In the replication analysis, risk alleles at both GATA3 SNPs were consistently over-represented in AYA ALL cases compared to non-ALL controls: rs3824662 (P=1.52×10^{-10}, OR=2.21, [1.72 to 2.83]), and rs3781093 (P=1.0×10^{-7}, OR=1.96, [1.52 to 2.54], Table 1, Figure 3B and Supplemental Figure 5B). rs3824662 was validated across race/ethnicity in the replication group (i.e., EAs [35% vs. 18%, P=2.0×10^{-7}], Hispanics [55% vs. 39%, P=0.005] and AAs [13% vs. 9%, P=0.035], Figure 3B). In contrast, rs3781093 was significant in EAs and Hispanics but not in AAs (Supplemental Figure 5B).
We also examined the association signals in AYAs for susceptibility loci previously identified in pediatric populations (Supplemental Table 1). ARID5B, IKZF1, and PIP4K2A variants were nominally significant in AYAs in the discovery GWAS and/or in the replication analyses. In contrast, CEBPE and CDKN2A/2B were not associated with ALL risk in AYAs in either discovery or replication cohorts. These results imply both similarities and differences in genetic predisposition to ALL between children and AYAs.

**GATA3 SNP rs3824662 and ALL subtypes in AYAs**

We further analyzed the association of the GATA3 SNP rs3824662 with somatic ALL genomic abnormalities. Among the AYA ALL cases in the discovery cohort, the risk allele at rs3824662 was under-represented among hyperdiploid ALL cases (22% vs. 37%; $P=0.03$, Figure 4, Supplemental Table 2), with similar trend for TCF3-PBX1 and ETV6-RUNX1 ALL albeit not statistically significant. In contrast, the ALL risk allele frequency of rs3824662 was higher in AYA ALL cases with the Ph-like gene expression profile than those without this signature (48% vs. 32%, $P=0.02$, Figure 4 and Supplemental Table 2). This was consistent with our prior reports of GATA3 as a susceptibility gene for Ph-like ALL, even though there was no overlap in cases included in the current AYA ALL GWAS and those in our previous Ph-like ALL GWAS. Within Ph-like ALL, there was a trend with A allele further enriched in cases involving CRLF2 rearrangements ($P=0.06$, Figure 4).

Importantly, even after excluding Ph-like ALL cases, risk allele at rs3824664 was still more common in AYA ALL cases compared to non-ALL controls (rs3824662: $P=8.13 \times 10^{-5}$, OR=1.56 [95% CI=1.25-1.96]; rs3781093: $P=0.0002$, OR=1.53 [95% CI=1.21-1.92], Supplemental Figure 6). This suggested that the influence of the GATA3 variant on ALL susceptibility in AYAs extends beyond the predisposition to Ph-like subtype.
GATA3 SNP rs3824662 and age at ALL diagnosis

Finally, we examined the distribution of the GATA3 SNP genotype by age at diagnosis in a cohort of largely unselected patients enrolled on the COG P9900 protocols (N=1,827, age from 0.1 to 21 years). Dividing patients into four consecutive age groups (<5, 5-10, 10-15 and >15 years), we observed a clear progressive increase in the risk allele frequency at rs3824662 ($P=6.29 \times 10^{-11}$, Figure 5) with increasing allelic odds ratio (i.e. relative risk of ALL conferred by each copy of the A allele at rs3824662, Figure 5 inset plot): 0.96 (95% CI=0.85 to 1.09), 1.26 (1.08 to 1.48), 1.48 (1.19 to 1.84), and 2.40 (1.81 to 3.19). Similar correlation between GATA3 genotype frequency and age was evident irrespective of genetic ancestry, but the GATA3 risk allele was markedly more common in Hispanic Americans (i.e. individuals with high Native American genetic ancestry, Figure 5). To examine whether the association with age is confounded by ALL genetic subtype, we compared rs3824662 allele frequency by age in the COG P9900 protocols after stratifying ALL cases into TCF3-PBX1, ETV6-RUNX1, high hyperdiploid, MLL-rearranged and B-other. There was a trend that for the risk allele at this SNP to be more frequent in cases older than 16 years relative to those below 16 in five subtypes examined, although with a limited sample size (Supplemental Figure 7). This suggests that GATA3 germline variants confer a general ALL disease risk in AYAs. In contrast, the frequency of ALL risk variant in ARID5B (rs10821936) decreased progressively with increasing age at diagnosis in the COG P9900 cohort ($P=0.006$), whereas PIP4K2A, CDKN2A/2B, IKZF1, and CEBPE variants were not related to age ($P>0.05$, data not shown).
DISCUSSION

Because ALL is the most common cancer in children, previous susceptibility GWAS studies understandably focused on pediatric populations. We hypothesized that ALL in AYAs has distinct tumor biology and genetic etiology, which potentially contributes to the disparities in treatment outcomes by age. To this end, we performed the first GWAS of ALL susceptibility specifically in the AYA population and identified a single genome-wide significant risk locus within the \textit{GATA3} gene on 10p14.

The susceptibility to ALL varies substantially by age. ALL risk first peaks between 2 and 5 years after birth, followed by gradual decrease into adulthood, but rises again in older individuals (>70), suggesting that differential combinations of environmental and genetic factors contribute to leukemogenesis at different ages. For example, it has been hypothesized that infection (and supposedly acquired immunity) may ameliorate susceptibility to ALL in young children\textsuperscript{41,42}, which may not be important in ALL that occurs later in life. Similarly, the \textit{in utero} occurrence of genomic lesions is characteristic in many (if not most) pediatric ALL cases\textsuperscript{43,44}, whereas such early origin of presumed initiating events may not be evident in AYAs ALL. Age-dependent differences in lymphocyte development and function are well documented in human and mouse systems\textsuperscript{45}, and rapid growth of hematopoietic cells may render them particularly susceptible to oncogenic assaults\textsuperscript{46}. Thus, it can be postulated that specific ALL susceptibility genes are required during a particular stage of hematopoietic development and preferentially influence ALL risk within a certain age range. For example, loss of \textit{Arid5b} in mouse resulted in reduction of lymphoid cells in bone marrow within 3 weeks after birth, but the effect became blunted by 6 weeks\textsuperscript{47}. In fact, germline \textit{ARID5B} variants also exhibited increasing influence on ALL predisposition in children as age decreases\textsuperscript{15}. 


GATA3 encodes for a transcription factor critical for lymphoid cell lineage commitment and early T cell differentiation, and loss-of-function somatic mutations have been discovered in early T-cell precursor ALL. Germline polymorphisms in GATA3, however, appear more important for B cell malignancies. We recently reported that rs3824662 was significantly associated with susceptibility to Ph-like ALL in children and risk of relapse. A contemporaneous study by Migliorini et al. reported the same ALL susceptibility variant in GATA3 in children of European descent and associated it with relapse. Particularly of note, GATA3 risk variants also appeared enriched in older children even within their predominantly pediatric cohort. The association of rs3824662 with ALL relapse is in line with the negative prognosis by age and higher frequency of the GATA3 variant in AYAs with ALL. Nevertheless, it is unclear whether poor prognosis conferred by GATA3 variant was driven by its association with high risk subtype (i.e., Ph-like ALL), novel somatic genomic aberrations specific to AYA, and/or host biology related to anti-leukemic drug response. Interestingly, in AYA cases included in the discovery GWAS, the number of the risk allele at rs3824662 was significantly associated with GATA3 expression in ALL blasts ($P=0.02$, Supplemental Figure 8), consistent with our previous report in pediatric ALL cases of this variant functioning as a cis-acting regulatory element of GATA3 transcription.

The overrepresentation of the GATA3 variant in AYAs is consistent with its association with Ph-like ALL for which the frequency increases with age. However, the risk variant at rs3824662 remained significantly associated with susceptibility to AYA ALL cases without Ph-like expression pattern, suggesting the link to Ph-like contributed only partly to the genome-wide significant association signal at rs3824662. In fact, GATA3 risk allele tended to be more common in ALL cases aged 16 and above than in those below 16 consistently across different
genetic subtypes, plausibly conferring a general ALL risk in AYAs. It remains unknown how the
GATA3 variants influence the risk of developing ALL in older adults, including the elderly (>60
years). Future studies including this age group may provide insights on molecular etiology of
ALL across age spectrum. It is also noteworthy that MLL-rearranged cases had the second
highest GATA3 risk variant frequency (Figure 4), although the number of patients was
relatively small and the difference did not reach statistical significance. Future studies are
warranted to comprehensively characterize potential interactions of germline GATA3 variants
with somatic genomic lesions in ALL.

In conclusion, our GWAS identified inherited GATA3 genetic variants that strongly influence
ALL susceptibility in adolescents and young adults, shedding new light on potential age-related
differences in ALL biology and treatment outcome.
ACKNOWLEDGEMENTS

We thank the patients and parents who participated in the clinical trials included in this study. Genome-wide genotyping of COG P9904/P9905 samples was performed by the Center for Molecular Medicine with the generous financial support from the Jeffrey Pride Foundation and the National Childhood Cancer Foundation. V.P.A. is supported by a St. Jude Children’s Research Hospital Academic Programs Special Fellowship and by a Spanish Ministry of Education Fellowship Grant. J.J.Y. is supported by the American Society of Hematology Scholar Award and by the Order of St. Francis Foundation. K.G.R. is supported by a Leukemia and Lymphoma Society Fellow Award and Alex’s Lemonade Stand Foundation Young Investigator Award. S.P.H. is the Ergen Family Chair in Pediatric Cancer. C.G.M. is a Pew Scholar in the Biomedical Sciences and a St. Baldrick’s Scholar. H.Z. is a St. Baldrick’s International Scholar. We thank M. Shriver (Pennsylvania State University) for sharing SNP genotype data of the Native American references. This work was supported by the National Institutes of Health (grant numbers CA145707, CA156449, CA21765, CA36401, CA98543, CA114766, CA98413, CA140729, CA176063, and GM92666), in part by the intramural Program of the National Cancer Institute, by a Stand up to Cancer Innovative Research Grant, and by the American Lebanese Syrian Associated Charities (ALSAC) of St Jude Children’s Research Hospital. The study sponsors were not directly involved in the design of the study, the collection, analysis, and interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript.
AUTHORSHIP

Conception and design: Virginia Perez-Andreu, M.D. Ph.D., Kathryn G. Roberts, Ph.D., Heng Xu, Ph.D., Charles G. Mullighan, M.D., and Jun J. Yang, Ph.D.

Provision of study materials or patients: Richard C. Harvey, Ph.D., Debbie Payne-Turner, B.S., I-Ming Chen, D.V.M., William L. Carroll, M.D., Nyla A. Heerema, Ph.D., Andrew J. Carroll, Ph.D., Elizabeth A. Raetz, M.D., Julie M. Gastier-Foster, Ph.D., Guido Marcucci, M.D., Clara D. Bloomfield, M.D., Krzysztof Mrózek, M.D. Ph.D., Jessica Kohlschmidt, Ph.D., Wendy Stock, M.D., Steven M. Kornblau, M.D., Marina Konopleva, M.D., Elisabeth Paietta, Ph.D., Jacob M. Rowe, M.D., Selina M. Luger, M.D., Martin S. Tallman, M.D., Michael Dean, Ph.D., Esteban G. Burchard, M.D., Dara G. Torgerson Ph.D., Feng Yue, Ph.D., Yanli Wang, Ching-Hon Pui, M.D., Sima Jeha, M.D., Mary V. Relling, Pharm.D., William E. Evans, Pharm.D., Daniela S. Gerhard, Ph.D., Mignon L. Loh, M.D., Stephen P. Hunger, M.D., and Cheryl L. Willman M.D.

Collection and assembly of data: Virginia Perez-Andreu, M.D. Ph.D., Kathryn G. Roberts, Ph.D., Heng Xu, Ph.D., Meenakshi Devidas, Ph.D., I-Ming Chen, D.V.M., Cheryl L. Willman M.D., Richard C. Harvey, Ph.D., Mary V. Relling, Pharm.D., and William E. Evans, Pharm.D.

Data analysis and interpretation: Virginia Perez-Andreu, M.D. Ph.D., Heng Xu, Ph.D., Colton Smith, Ph.D., Wenjian Yang, Ph.D., Hui Zhang Ph.D., Meenakshi Devidas, Ph.D., Richard C. Harvey, Ph.D., I-Ming Chen, D.V.M., and Jun J. Yang, Ph.D.

Manuscript writing: Virginia Perez-Andreu, M.D. Ph.D., Heng Xu, Ph.D., Charles G. Mullighan, M.D., and Jun J. Yang, Ph.D.

Final approval of manuscript: All authors
CONFLICT OF INTEREST The author(s) declare no competing financial interest.

CORRESPONDENCE
Jun J. Yang Ph.D. Department of Pharmaceutical Sciences, MS 313, St. Jude Children’s Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105-3678; e-mail: jun.yang@stjude.org, Phone: (901) 595-2517, FAX: (901) 595-8869
Charles G. Mullighan M.D., Department of Pathology, MS 342, St Jude Children’s Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105-3678, e-mail: charles.mullighan@stjude.org, Phone: (901) 595-3387, FAX: (901) 595-5947.
REFERENCES


51. Integrated Genomic and Mutational Profiling Of Adolescent and Young Adult ALL Identifies a High Frequency Of BCR-ABL1-Like ALL with Very Poor Outcome. Vol. 122; 2013.
Table 1. Association of *GATA3* SNPs with AYA ALL susceptibility in the discovery GWAS and replication cohort.

<table>
<thead>
<tr>
<th>Chr Position</th>
<th>SNP</th>
<th>Alleles</th>
<th>Cohort</th>
<th>RAF(^4) in AYA Cases</th>
<th>Genotype Count in Cases(^5)</th>
<th>RAF(^4) in non-ALL Controls</th>
<th>Genotype Count in Controls(^5)</th>
<th>AYA ALL vs non-ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8144214</td>
<td>rs3824662</td>
<td>A/C</td>
<td>Discovery</td>
<td>35%</td>
<td>36</td>
<td>145</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>39%</td>
<td>24</td>
<td>78</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>8141933</td>
<td>rs3781093</td>
<td>C/T</td>
<td>Discovery</td>
<td>33%</td>
<td>34</td>
<td>137</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replication</td>
<td>35%</td>
<td>18</td>
<td>75</td>
<td>64</td>
</tr>
</tbody>
</table>

Abbreviations: Chr, chromosome; OR, odds ratio; CI, confidence interval.

\(^1\)Association of SNP genotype and ALL was evaluated by comparing allele frequency between ALL and non-ALL, after adjusting for genetic ancestry. \(^2\)Chromosomal locations are based on hg18. \(^3\)Bold indicates risk allele for ALL. \(^4\)RAF, risk allele frequency (allele A at rs38246623 and allele C at rs3781093). \(^5\)Genotype is denoted by RR (homozygous for the risk allele), WW (homozygous for the wildtype allele), or RW (heterozygous). \(^6\)P-values were estimated by the logistic regression test and OR represents the increase in risk of developing ALL for each copy of the risk allele compared with subjects who don’t carry the risk allele.
FIGURE LEGENDS

Figure 1. GWAS study design. ALL susceptibility variants were identified by comparing SNP genotype frequency in AYA ALL cases compared to non-ALL controls in the discovery GWAS, followed by replication. AYA: adolescent and young adult (16-39 years); COG: Children’s Oncology Group; CALGB: The Alliance-Cancer and Leukemia Group B; ECOG: Eastern Cooperative Oncology Group; MDACC: MD Anderson Cancer Center; SJ: St Jude Children’s Research Hospital.

Figure 2. Genome-wide association of SNP genotype with ALL susceptibility in AYAs. The association between genotype and ALL susceptibility was evaluated by using logistic regression model for 635,297 SNPs in 308 AYA ALL cases and 6,661 non-ALL controls. P values (-log₁₀ [P value], y axis) were plotted against respective chromosomal position of each SNP (x axis). Points above the blue horizontal line indicate SNPs achieving the genome-wide significant threshold (P<5×10⁻⁸). Gene symbol was indicated for the GATA3 locus at 10p14.

Figure 3. Association of GATA3 SNP rs3824662 with ALL in AYAs by race/ethnicity. In the discovery group (A) the A allele at rs3824662 was overrepresented in AYA ALL cases relative to non-ALL controls. This association was true within the European Americans (>95% European genetic ancestry), African Americans (>70% African ancestry), or Hispanic Americans (>10% Native American genetic ancestry and Native American ancestry > African genetic ancestry). Similar association was confirmed in the replication group (B, one-tailed test). Genetic ancestry was determined by using STRUCTURE (version 2.2.3) with HapMap CEU, YRI, CHB/JPT, and indigenous Native Americans as reference populations.

Figure 4. GATA3 SNP genotype and ALL genetic subtypes in AYAs. The allele frequency of rs3824662 varied substantially by ALL somatic genomic abnormalities, with the ALL risk
allele under-represented in hyperdiploid cases and more common in the Ph-like subtype. Numbers are based on the ALL cases included in the discovery GWAS (N=308). The frequency of A allele at rs3824662 was 20% among unrelated non-ALL controls (MESA).

**Figure 5. GATA3 SNP rs3824662 and age at ALL diagnosis.**

In a largely unselected cohort of ALL cases enrolled on the Children’s Oncology Group (COG) P9900 trials (N=1,827), the frequency of ALL risk allele at rs3824662 was positively correlated with patient age at diagnosis consistently across race/ethnicity. Inset figure: the relative risk of ALL (odds ratio, OR) conferred by each copy of the A allele at rs3824662 increased progressively with age, as estimated by logistic regression after adjusting for genetic ancestry. Dotted line is OR of 1. Unrelated subjects from the Multi-ethnic Study of Atherosclerosis (MESA) were considered as non-ALL controls.
Figure 1. GWAS study design. ALL susceptibility variants were identified by comparing SNP genotype frequency in AYA ALL cases compared to non-ALL controls in the discovery GWAS, followed by replication. AYA: adolescent and young adult (16-39 years); COG: Children’s Oncology Group; CALGB: The Alliance-Cancer and Leukemia Group B; ECOG: Eastern Cooperative Oncology Group; MDACC: MD Anderson Cancer Center; SJ: St Jude Children’s Research Hospital.
Figure 2. Genome-wide association of SNP genotype with ALL susceptibility in AYAs. The association between genotype and ALL susceptibility was evaluated by using logistic regression model for 635,297 SNPs in 308 AYA ALL cases and 6,661 non-ALL controls. \( P \) values (-log\(_{10}\) [\( P \) value], y axis) were plotted against respective chromosomal position of each SNP (x axis). Points above the blue horizontal line indicate SNPs achieving the genome-wide significant threshold (\( P<5\times10^{-8} \)). Gene symbol was indicated for the GATA3 locus at 10p14.
Figure 3. Association of GATA3 SNP rs3824662 with ALL in AYAs by race/ethnicity. In the discovery group (A) the A allele at rs3824662 was overrepresented in AYA ALL cases relative to non-ALL controls. This association was true within the European Americans (>95% European genetic ancestry), African Americans (>70% African ancestry), or Hispanic Americans (>10% Native American genetic ancestry and Native American ancestry > African genetic ancestry). Similar association was confirmed in the replication group (B, one-tailed test). Genetic ancestry was determined by using STRUCTURE (version 2.2.3) with HapMap CEU, YRI, CHB/JPT, and indigenous Native Americans as reference populations.
Figure 4. GATA3 SNP genotype and ALL genetic subtypes in AYAs. The allele frequency of rs3824662 varied substantially by ALL somatic genomic abnormalities, with the ALL risk allele under-represented in hyperdiploid cases and more common in the Ph-like subtype. Numbers are based on the ALL cases included in the discovery GWAS (N=308). The frequency of A allele at rs3824662 was 20% among unrelated non-ALL controls (MESA).
Figure 5. GATA3 SNP rs3824662 and age at ALL diagnosis. In a largely unselected cohort of ALL cases enrolled on the Children’s Oncology Group (COG) P9900 trials (N=1,827), the frequency of ALL risk allele at rs3824662 was positively correlated with patient age at diagnosis consistently across race/ethnicity. Inset figure: the relative risk of ALL (odds ratio, OR) conferred by each copy of the A allele at rs3824662 increased progressively with age, as estimated by logistic regression after adjusting for genetic ancestry. Dotted line is OR of 1. Unrelated subjects from the Multi-ethnic Study of Atherosclerosis (MESA) were considered as non-ALL controls.
A genome-wide association study of susceptibility to acute lymphoblastic leukemia in adolescents and young adults