Regional evaluation of childhood acute lymphoblastic leukemia genetic susceptibility loci among Japanese

Kevin Y. Urayama1,2, Masatoshi Takagi3, Takahisa Kawaguchi4, Keitaro Matsuo5, Yoichi Tanaka6, Yoko Ayukawa7, Yuki Arakawa8, Daisuke Hasegawa9, Yuki Yuza9, Takashi Kaneko9, Yasushi Noguchi10, Yuichi Taneyama11, Setsuo Ota12, Takeshi Inukai13, Masakatsu Yanagimachi14, Dai Keino15, Kazutoshi Koike16, Daisuke Toyama17, Yozo Nakazawa18, Hidemitsu Kurosoawa19, Kozue Nakamura20, Koichi Moriwaki21, Hiroaki Goto22, Yujin Sekinaka23, Daisuke Morita24, Motohiro Kato25, Junko Takita25, Toshihiro Tanaka26,27, Johji Inazawa27, Katsuyoshi Koh7, Yasushi Ishida28, Akira Ohara29, Shuki Mizutani30, Fumihiko Matsuda31 & Atsushi Manabe31

Genome-wide association studies (GWAS) performed mostly in populations of European and Hispanic ancestry have confirmed an inherited genetic basis for childhood acute lymphoblastic leukemia (ALL), but these associations are less clear in other races/ethnicities. DNA samples from ALL patients (aged 0–19 years) previously enrolled onto a Tokyo Children’s Cancer Study Group trial were collected during 2013–2015, and underwent single nucleotide polymorphism (SNP) microarray genotyping resulting in 527 B-cell ALL for analysis. Cases and control data for 3,882 samples from the Nagahama Study Group and Aichi Cancer Center Study were combined, and association analyses across 10 previous GWAS-identified regions were performed after targeted SNP imputation. Linkage disequilibrium (LD) patterns in Japanese and other populations were evaluated using the varLD score based on 1000 Genomes

1Department of Social Medicine, National Center for Child Health and Development, Tokyo, Japan. 2Graduate School of Public Health, St. Luke’s International University, Tokyo, Japan. 3Department of Pediatrics and Developmental Biology, Tokyo Medical and Dental University, Tokyo, Japan. 4Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan. 5Division of Molecular and Clinical Epidemiology, Aichi Cancer Center Research Institute, Aichi, Japan. 6Department of Clinical Pharmacy, Center for Clinical Pharmacy and Sciences, School of Pharmacy, Kitasato University, Tokyo, Japan. 7Department of Hematology/Oncology, Saitama Children’s Medical Center, Saitama, Japan. 8Department of Pediatrics, St. Luke’s International Hospital, Tokyo, Japan. 9Department of Hematology/Oncology, Tokyo Metropolitan Children’s Medical Center, Tokyo, Japan. 10Department of Pediatrics, Japanese Red Cross Narita Hospital, Chiba, Japan. 11Department of Hematology/Oncology, Chiba Children’s Hospital, Chiba, Japan. 12Department of Pediatrics, Teikyo University Children’s Medical Center, Chiba, Japan. 13Department of Pediatrics, University of Yamanashi, Yamanashi, Japan. 14Department of Pediatrics, Yokohama City University Graduate School of Medicine, Yokohama, Japan. 15Department of Pediatrics, St. Marianna University School of Medicine, Kawasaki, Japan. 16Division of Pediatric Hematology and Oncology, Ibaraki Children’s Hospital, Mito, Japan. 17Division of Pediatrics, Showa University Fujigaoka Hospital, Yokohama, Japan. 18Department of Pediatrics, Shinshu University School of Medicine, Matsumoto, Japan. 19Department of Pediatrics, Dokkyo Medical University, Tochigi, Japan. 20Department of Pediatrics, Teikyo University Hospital, Tokyo, Japan. 21Department of Pediatrics, Saitama Medical Center, Saitama Medical University, Saitama, Japan. 22Division of Hematology/Oncology & Regenerative Medicine, Kanagawa Children’s Medical Center, Yokohama, Japan. 23Department of Pediatrics, National Defense Medical College, Saitama, Japan. 24Children’s Cancer Center, National Center for Child Health and Development, Tokyo, Japan. 25Department of Pediatrics, The University of Tokyo Hospital, Tokyo, Japan. 26Department of Human Genetics and Disease Diversity, Tokyo Medical Dental University, Tokyo, Japan. 27Bioresource Research Center, Tokyo Medical and Dental University, Tokyo, Japan. 28Pediatric Medical Center, Ehime Prefectural Central Hospital, Matsuyama, Japan. 29Department of Pediatrics, Toho University, Tokyo, Japan. Correspondence and requests for materials should be addressed to K.Y.U. (email: urayama-k@ncchd.go.jp)
data. Risk associations for ARID5B (rs10821936, OR = 1.84, P = 6 × 10⁻¹²) and PIP4K2A (rs7088318, OR = 0.76, P = 2 × 10⁻⁸) directly transferred to Japanese, and the IKZF1 association was detected by an alternate SNP (rs1451367, OR = 1.52, P = 2 × 10⁻⁸). Marked regional LD differences between Japanese and Europeans was observed for most of the remaining loci for which associations did not transfer, including CEBPE, CDKN2A, CDKN2B, and ELK3. This study represents a first step towards characterizing the role of genetic susceptibility in childhood ALL risk in Japanese.

Scrutiny of the human genome through evaluation of common genetic variants has revealed hundreds of disease susceptibility loci. In childhood acute lymphoblastic leukemia (ALL), six regions that have replicated in several populations are now considered known susceptibility loci (likely representing associations with ARID5B, IKZF1, CEBPE, CDKN2A, PIP4K2A, and GATA3), with the majority of the evidence supported through studies conducted in populations of European and Hispanic descent⁶⁻⁷. Gains in statistical power achieved by recent meta-analyses of childhood ALL genome-wide association studies (GWAS) have resulted in the identification of risk-associated single nucleotide polymorphisms (SNPs) of comparatively lower allele frequencies and estimated magnitude of effects including those tagging the CDKN2B, LHPP, and ELK3 genes⁸⁻¹⁰. Furthermore, the concept that genetic susceptibility studies may potentially reveal race/ethnicity-specific associations was demonstrated by a recent GWAS conducted in a Chinese population which implicated a role for the WWOX gene that had not been observed in the numerous previous studies conducted in populations of European and Hispanic ancestry⁹.

Success of GWAS in identifying true disease-associated loci have largely been due to the consistently high standards in methodological rigor of the approach including, strict quality control for genotype data, attention to issues of statistical power and sample size, criteria for genome-wide significance, and integrating components of independent validation and/or functional evaluation of loci⁴. However, as is the case with other complex diseases, it is well-recognized that the known GWAS ‘hits’ in childhood ALL account for only a small proportion of the total estimated heritability⁷. Based on data from populations of European ancestry, it has been estimated that the currently known childhood ALL associated risk loci account for about 19 percent of the additive heritable risk, not accounting for potential impact of epistasis or gene-environment interactions⁸. Adding to this issue of ‘missing heritability’ is the realization that we currently know even less about the nature of childhood ALL genetic susceptibility in other populations, particularly Asians⁹.

The effects of known genetic susceptibility loci have yet to be fully confirmed in populations of non-European ancestry. Targeted validation attempts based on the same SNPs originally identified in mostly non-Hispanic whites and Hispanics have been performed in Chinese and other Asian populations, but findings have been inconsistent⁹⁻¹⁰. Assuming the same causal variant is operative across populations, a lack of association in Asians can be attributed to study flaws, but more likely due to reduced statistical power as a result of differences in allele frequency, strength of linkage disequilibrium (LD) with the causal variant(s), and/or the role of environmental exposures in affecting risk. Thus, a comprehensive characterization of genetic variation across the targeted genetic loci is required for an appropriate validation attempt in different populations.

To address this current gap in knowledge, we initiated an effort through the Tokyo Children’s Cancer Study Group (TCCSG) to assemble resources for genomic investigation of germline contributions to childhood ALL susceptibility and outcomes. Here, we report results of our targeted analysis of previous GWAS-identified childhood ALL risk loci in a large Japanese population. We evaluated the transferability of risk associations of specific SNP loci in Japanese and interpreted the finding in the context of quantified differences in LD within those loci across populations. Furthermore, we analyzed regional SNP data in order to identify alternate SNPs which may potentially confer stronger association in Japanese.

Materials and Methods

Study Population. In collaboration with a large network of 23 hospitals participating in the Tokyo Children’s Cancer Study Group (TCCSG), previously diagnosed childhood ALL patients visiting for a routine follow-up between 2013 and 2015 were invited to participate in this study. The TCCSG network includes nearly all clinical centers that diagnose and treat childhood ALL within the seven prefectures that comprise the Kanto and immediately surrounding regions¹¹⁻¹². Patients were considered eligible if they were 19 years of age or younger at the time of ALL diagnosis, enrolled onto a TCCSG treatment protocol, and self-identified as Japanese. Due to the nature of this sampling scheme, the study population comprised a survivorship population of childhood ALL patients.

Upon obtaining written informed consent, saliva samples using the Oragene Saliva DNA Self-Collection Kit (4 years of age and older) or Assisted Collection Kit (less than 4 years of age) (DNA Genotek, Ottawa, Canada) were collected from the patients with instruction by the attending physician or nurse during the follow-up outpatient visit. The collected samples were shipped at room temperature to a central laboratory (Tokyo Medical and Dental University) for processing, DNA extraction, and storage.

Controls comprised a subset of adult participants enrolled in two ongoing epidemiological studies of lifestyle-related chronic diseases in Japan, the Nagahama Study Group¹³ and Aichi Cancer Center Study¹⁴, in which large-scale genome-wide SNP genotyping had already been performed. The Nagahama Study is a community-based prospective cohort study comprising a representative sample of residents of Nagahama City in Shiga Japan¹⁵. The Aichi Cancer Center Study comprised a hospital-based cohort of non-cancer outpatient visitors¹⁶. Despite the name, the Aichi Cancer Center resembles a general hospital that does not require physician referral in which the majority of outpatients present with no abnormal findings by clinical examination. Population substructure across regions of Japan does exist; most notably between populations of Okinawa and the other main islands of Japan collectively. Although cases and controls were recruited from different regions of the main island, simulation studies have shown only minimal genomic inflation potential when considering
these two subpopulations. A history of childhood leukemia was not assessed in controls; however, the rarity of this disease suggests that any previous diagnosis of childhood leukemia in controls would have a minimal effect on the results of this study.

This study protocol was approved by the institutional review boards of St. Luke's International Hospital, Tokyo Medical and Dental University, Kyoto University, and all collaborating hospitals involved in patient recruitment. Written informed consent was obtained from the parents of each participant together with a written assent by the child where possible. Patients aged 16 to 19 years were asked to provide written informed consent together with parental consent; those aged 20 or older did not require parental consent. This study was conducted in accordance with the Declaration of Helsinki.

**Genotyping and Quality Control.** DNA extraction from childhood ALL patients' saliva samples were performed using the Oragene prepIT DNA Extraction Kit (DNA Genotek) based on the manufacturer's instruction. The approximately 2 mL saliva samples obtained from the Oragene Self-Collection Kits yielded, on average, a total of about 50 ug of genomic DNA.

Genome-wide SNP genotyping was attempted on 621 patient samples using the Illumina HumanCoreExome-12 v1.1 BeadChip (San Diego, CA) which contained probes for approximately 550,000 SNPs. Existing control data were genotyped previously using variable versions of the same Illumina HumanCoreExome BeadChip. Quality control steps were conducted within cases and each of the two different control sample series separately. SNPs were excluded if the genotype call rate was below 99%, the distribution of genotypes clearly deviated from that expected by Hardy-Weinberg equilibrium (HWE) \((P < 1 \times 10^{-6})\), or the minor allele frequency was less than 0.01. Samples were excluded if showed a genotyping success rate of less than 95% (51 cases and 4 controls) and relatedness based on an identity-by-descent analysis (1 case and 119 controls). In addition, principal components analysis (PCA) based on a genome-wide subset of SNPs in low LD (pruned at \(r^2 < 0.1\)) that passed quality control steps was performed on a known ethnically homogeneous population of Japanese ancestry (International HapMap Project) together with cases and controls. The PCA was conducted using the EIGENSTRAT 2.0 software package and outlier samples were excluded (2 cases and 5 controls). In result, after quality control steps and excluding 40 T-cell ALL patients, the final population for analysis included a total of 527 Japanese B-cell ALL cases and 3,882 controls with data available for 171,547 SNPs that were overlapping across the genotyped case and control series.

Targeted SNP imputation was performed on the combined case-control dataset for 10 genomic regions reported in previous childhood ALL GWAS (Table 1) using ShapeIT2 and Minimac3, and the 1000 Genomes Project Phase III Version 5 as the reference population. Poorly imputed SNPs defined by an \(R^2 < 0.5\) were excluded from the analyses. Considering the gene and its broad surrounding region (about 100-kb flanking) for each locus, a total of 113 SNPs were excluded among 14,457 total SNPs imputed across the 10 regions. On average, about 0.8 percent of SNPs per locus were excluded based on this quality control threshold. Due to restrictions stipulated by the institutional review board approvals, data were not be made publicly available, but may be available on request in compliance with the policies and procedures of the TCCSG.

**Statistical Analysis.** We first tested the association between childhood ALL and 16 SNPs across the 10 genes (Table 1) identified in previous GWAS. SNPs for evaluation were selected based on the strongest result reported from the first study to report the association. Multiple SNPs tagging the same genomic region were selected if the SNP was examined across several studies. We examined the role of additional genetic variation across the entire span of the 10 targeted genes, including a 10-kb flanking region on both ends. The association between each genetic variant and risk of childhood ALL was estimated by the odds ratio (OR) per allele and 95% confidence intervals (CI) using multiple logistic regression assuming a log-additive genetic model. Genome-wide association analysis of the 171,547 SNPs showed evidence of genomic inflation (\(\lambda > 1.10\)); all analyses were adjusted for 10 PCA eigenvectors (\(\lambda = 1.05\)). For the test of specific previously reported GWAS SNPs, a nominal p-value of less than 0.05 was considered statistically significant. For the examination of other potentially associated SNPs across the genomic regions, to account for multiple comparisons in the presence of LD between SNPs, we calculated adjusted p-values based on 10,000 permutations of case-control status and considered p-values below a family-wise type I error threshold of 0.05 to be statistically significant. Analyses were conducted using PLINK2 and SAS software version 9 (SAS, Cary, NC). The LocusZoom web-based resource was used to generate plots of association results by genomic region.

Differences across race/ethnic populations in regional patterns of LD flanking a 10-kb region on both ends of the SNPs were quantified using the variation in LD (varLD) score applied to the 1000 Genomes Phase 3 data. The varLD score is an algorithm based on comparing regional patterns of correlation previously developed by Teo et al. to quantify differences in LD within defined regions. With the exception of the WWOX locus, the Japanese (JPT) population was compared to the combined population of European ancestry (EUR); for the WWOX locus, JPT was compared to the combined Han Chinese and Southern Han Chinese (CHB-CHS) representing the population in which the locus was originally identified. Permutation procedures were performed to determine Monte Carlo statistical significance by comparing the estimated varLD score to the null distribution of varLD scores after successive re-sampling of the two populations from the combined data; 10,000 iterations were performed. Since 9 genomic loci were tested (CDKN2A-CDKN2B were evaluated as one region), an empirical p-value of less than 0.006 was considered statistically significant for the varLD evaluation. All statistical tests were two-sided.

**Results**

Association analyses were performed on a total of 527 B-cell ALL cases and 3,882 controls. Median age at ALL diagnosis was 4.5 years (range: 0.3–16.8 years). The risk of childhood B-cell ALL associated with 16
SNPs (representing 10 genes) reported in previous GWAS was evaluated in this Japanese population (Table 1). The ARID5B SNPs showed strong evidence of an association with the highest risk observed for rs10821936 (OR = 1.84, 95% CI = 1.60–2.13, P = 6.04 × 10^{-17}). Of the remaining loci, the 2 correlated PIP4K2A SNPs evaluated showed an association with childhood ALL (rs10828317, OR = 0.76, 95% CI = 0.65–0.88, P = 3.03 × 10^{-4}) as well. The GATA3 rs3824662 association was only suggestive (OR = 1.19, 95% CI = 0.93–1.33, P = 0.251). The WWOX SNP, rs1121404, recently identified to be associated with childhood ALL in Japanese, showed no association in Japanese (OR = 1.04, 95% CI = 0.90–1.19, P = 0.623).

Risk allele frequencies and association estimates across various races/ethnicities are presented in Table 2. Among the loci identified through GWAS, only ARID5B SNPs showed a consistent association across the race/ethnic populations despite marked differences in allele frequencies. Although only marginally significant in Chinese (rs7089424, OR = 1.15, 95% CI = 0.90–1.38, P = 0.011), the WWOX SNP, rs1121404, recently identified to be associated with childhood ALL in Japanese, showed no association in Japanese (OR = 1.04, 95% CI = 0.90–1.19, P = 0.623).

Using available SNP data across all 10 genetic loci including 10-kb flanking regions on both ends of the target genes, B-cell ALL risk associations were identified for alternate SNPs in IKZF1 (rs1451367, OR = 1.52, 95% CI = 1.28–1.80, P = 1.9 × 10^{-6}) as well. For the two genetic loci where the SNP associations directly transferred to Japanese, rs4245959 in ARID5B (OR = 1.86, P = 2.1 × 10^{-17}) and rs12146350 in PIP4K2A (OR = 0.72, P = 2.7 × 10^{-4}) showed slightly stronger p-values, and both were in strong LD (r^2 > 0.90) in Japanese with the originally reported respective SNPs. The rs4245959 ARID5B SNP is also in strong LD (r^2 > 0.96) with the recently reported functional SNP, rs7090445 (OR = 1.85, P = 3.1 × 10^{-17}), identified by Studt et al. in which they showed influences on enhancer activity and RUNX3 binding. For the remaining genetic loci, alternate SNPs with
Table 2. Summary of genetic variants and childhood ALL risk associations across races/ethnicities identified through genome-wide association studies. Abbreviations: RAP, risk allele frequency; OR, odds ratio; Ref, reference; SNP, single nucleotide polymorphism. aFrequency of the allele in controls conferring an increased risk of childhood ALL. bOdds ratios indicate the risk associated with the each additive increase in risk conferring allele. cRisk allele frequency obtained from the Human Genome Diversity Project (HGDP) dCalculations pertain to the designated risk allele indicated in the primary report. Risk estimate results may not be greater than 1.0 in East Asians.

| Gene/SNP   | European Ancestry | Hispanic Ancestry | African Ancestry | Chinese Ancestry | Japanese Ancestry |
|------------|-------------------|-------------------|------------------|------------------|-------------------|
|            | RAP   | OR    | P    | Ref  | RAP   | OR    | P    | Ref  | RAP   | OR    | P    | Ref  | RAP   | OR    | P    |
| ARID5B     |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs10821936 | 0.35  | 1.91  | 1 × 10^{-12} | < 0.001 | 0.47  | 1.95  | 4 × 10^{-11} | 0.01 | 1.52  | 4 × 10^{-7} | 0.38  | 1.43  | 5 × 10^{-4} | 0.35  | 1.84  | 6 × 10^{-12} |
| IKZF1      |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs41332601 | 0.28  | 1.69  | 1 × 10^{-12} | < 0.001 | 0.27  | 1.46  | 1 × 10^{-3} | 0.02 | —     | —     | —     | 0.14  | 1.20  | 0.197 | 0.09  | 1.19  | 0.164   |
| rs11978267 | 0.27  | 1.69  | 9 × 10^{-11} | < 0.001 | 0.26  | 1.31  | 0.01  | 0.02 | 0.19  | 1.59  | 5 × 10^{-2} | 0.09  | 1.13  | 0.321 |       |       |         |
| CEBPE      |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs2239633  | 0.52  | 1.34  | 2.9 × 10^{-3} | < 0.001 | 0.61  | 1.35  | 6.6 × 10^{-3} | 0.27 | —     | —     | —     | 0.64  | 1.18  | 0.117 | 0.53  | 1.08  | 0.282   |
| rs4982731  | 0.28  | 1.29  | 9.1 × 10^{-6} | < 0.001 | 0.39  | 1.58  | 2.3 × 10^{-6} | 0.01 | 1.13  | 0.410 | —     | 0.13  | 1.11  | 0.306 |       |       |         |
| CDKN2A, CDKN2B |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs3731217  | 0.86  | 1.41  | 3 × 10^{-11} | < 0.001 | 0.88  | 1.76  | 5 × 10^{-3} | 0.32 | —     | —     | —     | 0.80  | 1.04  | 0.769 | 0.82  | 0.90  | 0.251   |
| rs6624663  | 0.10  | 1.48  | 2 × 10^{-10} | < 0.001 | 0.07  | 1.45  | 0.034 | 0.02 | 1.12  | 0.620 | —     | 0.01  | 0.85  | 0.625 |       |       |         |
| rs17756311 | 0.09  | 1.43  | 3 × 10^{-11} | < 0.001 | 0.06  | 1.36  | 0.100 | 0.01 | 1.12  | 0.620 | —     | 0.01  | 0.88  | 0.695 |       |       |         |
| PIP4K2A    |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs7088318  | 0.59  | 1.25  | 5 × 10^{-4}  | < 0.001 | 0.75  | 1.42  | 9 × 10^{-3} | 0.02 | 1.65  | 1 × 10^{-1} | 0.32  | 1.32  | 0.013 | 0.31  | 1.15  | 0.058   |
| GATA3      |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs3824662  | 0.17  | 1.31  | 9 × 10^{-12} | < 0.001 | 0.75  | 1.42  | 9 × 10^{-3} | 0.02 | 1.65  | 1 × 10^{-1} | 0.32  | 1.32  | 0.013 | 0.31  | 1.15  | 0.058   |
| LHPP       |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs35837782 | 0.62  | 1.21  | 1 × 10^{-15} | < 0.001 | 0.62  | 1.21  | 1 × 10^{-15} | 0.27 | 1.43  | 5 × 10^{-10} | 0.38  | 1.38  | 5 × 10^{-10} | 0.38  | 1.04  | 0.623   |
| ELK3       |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs4762284  | 0.29  | 1.19  | 8 × 10^{-4}  | < 0.001 | 0.27  | 1.43  | 5 × 10^{-10} | 0.38  | 1.38  | 5 × 10^{-10} | 0.38  | 1.38  | 5 × 10^{-10} | 0.38  | 1.04  | 0.623   |
| WWOX       |       |       |      |      |       |       |      |      |       |       |      |      |       |       |      |
| rs1121404  | 0.12  | 0.77  | 0.01  | 0.001 | 0.14  | 1.20  | 0.017 | 0.01 | 1.12  | 0.620 | —     | 0.01  | 0.88  | 0.695 |       |       |         |

A nominal p-values of less than 0.05 were identified, but were not statistically significant after adjustment for the number of SNPs tested across the respective regions.

To examine whether the non-transferability of association may be due to population differences in regional LD structure, varLD scores were calculated using 1000 Genomes Project Phase 3 data for Japanese, Han Chinese, and populations of European ancestry (Table 3). The regions surrounding the ARID5B and PIP4K2A SNPs, loci that directly replicated in Japanese, did not show statistically significant evidence of regional LD differences between populations of European ancestry and Japanese. Regions surrounding the IKZF1 SNPs also showed minimal evidence of regional LD differences, but the previous GWAS-identified SNP associations did not directly transfer to Japanese. However, alternate statistically significant SNPs within IKZF1 were identified (described above). With the exception of LHPP and WWOX, the four additional genetic loci in which the association did not transfer to Japanese showed strong evidence of regional LD structure differences based on varLD evaluations.

Discussion

Aided by successful validation across multiple populations, the genome-wide association analysis approach has led to the identification of several genetic loci involved in childhood ALL risk. However, there is still uncertainty about the role of these loci and consistency of specific SNP associations in East Asians with the majority of robust studies being performed primarily in populations of European and Hispanic ancestries. In our targeted evaluation of 16 previous GWAS-reported SNPs, we observed that the risk associations of those in ARID5B and PIP4K2A directly transfer to the Japanese population. The involvement of IKZF1 is also supported by the identification of alternate associated SNPs in proximity to the originally reported loci, and the GATA3 locus appears suggestive. However, this leaves the associations in the six remaining genes without clear evidence for a role in childhood ALL risk in East Asians. Examination of regional varLD scores showed that significant differences in LD between Japanese and the population in which the association was first reported were commonly observed in genes where the risk association did not transfer. Rather than concluding that the association is not present in Japanese, the varLD observations suggest that the associations may be obscured by differences in LD patterns and that other strategies are necessary to further clarify the role of the remaining six loci that did not transfer to this population.

Childhood ALL SNP associations in ARID5B first reported concurrently in studies performed in populations of European ancestry in the United States and the United Kingdom have been widely validated across multiple race/ethnic population, now including Japanese. The risk-conferring minor allele frequency of rs10821936...
in Japanese is similar to that of Europeans (MAF ~ 0.35), but is significantly higher in Hispanics (MAF ~ 0.45) and lower in populations of African ancestry (MAF ~ 0.20). Interestingly, this pattern is similar to the relative population differences in incidence of childhood ALL and evidence supports a role for this locus in partially explaining this difference. Based on available data from St. Jude Children's Research Hospital and descriptive statistics from the Surveillance, Epidemiology, and End Results Program in the US, it was estimated that about 30% of the observed racial differences in ALL incidence may be attributable to the higher frequency of the rs10821936 risk allele in non-Hispanic whites compared to blacks30. Characterization of genetic ancestry of the Children's Oncology Group Hispanic population showed increasing rs10821936 risk allele frequencies with increasing percentages of Native American ancestry31. Building on this observation, the California Childhood Leukemia Study reported increasing proportions of Native American ancestry to be associated with increasing risk of childhood ALL and showed that ARID5B contributes directly to the higher incidence in Hispanics compared to non-Hispanic whites32. However, the contribution of ARID5B is less clear in relation to Japanese given that this SNP has similar frequency and magnitude of effect as non-Hispanic whites despite known differences in incidence between the two populations.

Although consistently replicating in populations of European ancestry3, similar to studies performed in Chinese5, the IKZF1 SNP association did not transfer to the Japanese population. Comparison of LD patterns based on varLD score between Europeans and Japanese did not show evidence of marked difference across an approximately 25-kb region comprising the previously reported SNPs. However, the allele frequency of the SNPs are considerably lower in East Asians at about 0.10 or less compared to close to 0.30 in Europeans and Hispanics. The ability to analyze the effect of other SNPs across the flanking regions led to the identification of an alternate associated SNP (rs1451367) located within about 10-kb that is common in East Asians (MAF ~ 0.20), but rare in Europeans (MAF < 0.01). This suggests that variation in IKZF1 is also associated with risk of childhood ALL in Japanese; however, it cannot be concluded yet whether the SNP associations are representing the same causal locus across the populations.

Based on the results of the current analysis, evidence for childhood ALL risk associations with GWAS-identified SNPs in CEBPE, CDKN2A, CDKN2B, LHPP, ELK, and WWOX is lacking. Associations represented by other SNPs potentially tagging a causal locus within these genes were also not apparent. While the evidence is still limited, results could be influenced by differences in a gene-environment effect across populations not appropriately captured, or it may be possible that certain common SNPs identified in GWAS may be representing associations with rare causal variant(s) on the same haplotype background of the GWAS-identified tag SNP33. If rare causal variants are at play, even modest differences in haplotype structure of the regions may significantly affect detection potential, or it is possible that the variants may not be present in Japanese. As an example, the CDKN2A risk association originally identified through GWAS based on the common variant rs373121734 was
recently shown to be explained by a rare high-impact coding variant (rs3731249)\textsuperscript{35–37}. This variant is present in about two percent of Europeans, but is not present in Japanese. With the exception of \textit{LHPP}, all loci for which the associations did not transfer showed evidence of differences in genetic architecture between Japanese and Europeans based on \textit{varLD} score, whereas those that transferred did not show marked differences. In line with the common disease/common variant hypothesis\textsuperscript{38}, if the GWAS associations are instead tagging a common causal variant and assuming this variant is operative as a risk locus in Japanese as well, we would have expected the regional SNP coverage and statistical power of the current study to be sufficient to detect the association signal. The lack of association suggests a need for future studies to consider characterization of rare variants in order to fully understand the nature of these GWAS loci in Japanese.

Certain limitations inherent to this study may have also affected the results. Although our study was limited to B-cell lineage ALL similar to most previous GWAS, availability of molecular subtype data was incomplete for a large proportion of the patients. While heterogeneity by subtype in the magnitude of risk has been observed for several of the loci, effects exclusive to a specific subtype have not been clearly demonstrated and is likely not the reason for the lack of association observed. One exception may be the \textit{GATA3} risk locus identified in a GWAS of Ph-like ALL\textsuperscript{39} and another study that observed the association specifically in non-hyperdiploid B-cell ALL that lack the \textit{ETV6–RUNX1} fusion\textsuperscript{40}. Results for the \textit{GATA3} variant (rs3824662) in the current study were suggestive of an association among the total B-cell ALL series, but requires further evaluation in a subtype specific analysis for confirmation. Also, access to patients for recruitment into this study was through the outpatient mechanism which resulted in a study population of surviving patients. This may have led to over-representation of patients of certain disease profiles; however a 80 to 85 percent survival rate of childhood ALL, as reported by the TCCSG\textsuperscript{41}, suggests that the effect may have been minimal since our objectives focused on validating known GWAS hits, those of which were originally identified using general ALL patient populations that comprised of the most common ALL subtypes (versus a sequencing-based design targeting rare subtypes of poor prognosis). Finally, our data included imputed genotypes to enhance the coverage of genetic variation across the targeted genomic regions. Despite stringent quality control measures and advances in imputation methodologies, uncertainty still exists and may have introduced non-differential misclassification of genotypes and a reduction in statistical power to detect associations.

In this targeted evaluation of SNPs across regions previously identified in GWAS of childhood ALL, we showed that variation in \textit{ARID5B}, \textit{IKZF1}, \textit{PIP4K2A}, and possibly \textit{GATA3} contribute to the genetic susceptibility of childhood B-cell ALL in Japanese. There is a need to account for population-specificity in producing accurate

Figure 1. Regional plot of results of the association between SNPs in (a) \textit{ARID5B}, (b) \textit{PIP4K2A}, and (c) \textit{IKZF1} and risk of childhood B-cell ALL. Multiple logistic regression was performed assuming a log-additive genetic model and adjusting for 10 principal components eigenvectors. The $-\log_{10} (p$-value) for each SNP are plotted against their chromosomal position. The purple diamond (♦) indicates the strongest associated SNP in the region and the colors of the dots (●) represent the degree of linkage disequilibrium (based on $r^2$) in relation to that index SNP in Japanese. Recombination rates (cM/Mb) overlay the plots. Coordinates are based on human genome assembly GRCh37 build.
risk prediction estimates based on inherited genetic variation. Thus, this analysis serves as the first step towards characterizing the role of genetic variation in the susceptibility to childhood ALL in the Japanese population. Identification of potential novel loci, perhaps specific to the East Asian population or those more detectable due to enhanced LD with a causal locus and/or allele frequency differences, may be possible through a genome-wide association analysis after expansion of this population for increased statistical power.

References
1. Moriyama, T., Relling, M. V. & Yang, I. J. Inherited genetic variation in childhood acute lymphoblastic leukemia. Blood 125, 3988–3995, https://doi.org/10.1182/blood-2014-12-580001 (2015).
2. Urayama, K. Y., Chokkalingam, A. P., Manabe, A. & Mizutani, S. Current evidence for an inherited genetic basis of childhood acute lymphoblastic leukemia. Int J Hematol 97, 3–19, https://doi.org/10.1007/s12185-012-1220-9 (2013).
3. Hungate, E. A., et al. A variant at 9p21.3 functionally implicates CDKN2B in paediatric B-cell precursor acute lymphoblastic leukaemia etiology. Nat Commun 7, 10635, https://doi.org/10.1038/ncomms10635 (2016).
4. Vijayakrishnan, J., et al. A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukaemia at 10q26.13 and 12q23.1. Leukemia, https://doi.org/10.1038/leu.2016.271 (2016).
5. Shi, Y. et al. Identification of a novel susceptibility locus at 16q23.1 associated with childhood acute lymphoblastic leukemia in Han Chinese. Hum Mol Genet 25, 2873–2880, https://doi.org/10.1093/hmg/ddw112 (2016).
6. Hunter, D. J. Lessons from genome-wide association studies for epidemiology. Epidemiology 23, 363–367, https://doi.org/10.1097/ede.0b013e3182da7cc (2012).
7. Gibson, G. Hints of hidden heritability in GWAS. Nat Genet 42, 558–560, https://doi.org/10.1038/ng0710-558 (2010).
8. Urayama, K. Y. & Manabe, A. Genomic evaluations of childhood acute lymphoblastic leukaemia susceptibility across race/ethnicities. [Russho ketsueki] The Japanese Journal of Clinical Hematology 55, 2242–2248 (2014).
9. Shi, Y. et al. Secondary cancers among children with acute lymphoblastic leukaemia treated by the Tokyo Children’s Cancer Study Group protocol: a retrospective cohort study. Br J Haematol 164, 101–112, https://doi.org/10.1111/bjh.12602 (2014).
10. Kato, M. et al. Long-term outcome of 6-month maintenance chemotherapy for acute lymphoblastic leukaemia in children. Leukemia 31, 580–584, https://doi.org/10.1038/leu.2016.274 (2017).
11. Muro, S. et al. Relationship Among Chlamydia and Mycoplasma Pneumoniae Seropositivity, IKZF1 Genotype and Chronic Obstructive Pulmonary Disease in A General Japanese Population: The Nagahama Study. Medicine 95, e3371, https://doi.org/10.1097/MD.0000000000003371 (2016).
12. Inoue, M. et al. Epidemiology of pancreatic cancer in Japan: a nested case-control study from the Hospital-based Epidemiologic Research Program at Aichi Cancer Center (HERPACC). International journal of epidemiology 32, 257–262 (2003).
13. Urayama, K. Y., Chokkalingam, A. P., Manabe, A. & Mizutani, S. Current evidence for an inherited genetic basis of childhood acute lymphoblastic leukemia. Leukemia 580–584, https://doi.org/10.1038/leu.2016.274 (2017).
14. Sun, J. et al. Association between CEBPE Variant and Childhood Acute Leukemia Risk: Evidence from a Meta-Analysis of 22 Studies. PLoS One 10, e0125657, https://doi.org/10.1371/journal.pone.0125657 (2015).
15. Zeng, H. et al. Genotype imputation service and methods. Nature 507, 83–88, https://doi.org/10.1038/nature13134 (2014).
16. Price, A. L. et al. Next-generation genotype imputation service and methods. Nat Genet 48, 1284–1287, https://doi.org/10.1038/nmg.3656 (2016).
17. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559–575, https://doi.org/10.1093/ajhg/81.4.559 (2007).
18. Zaitlen, N. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337, https://doi.org/10.1093/bioinformatics/btq419 (2010).
19. O’Connell, J. R. et al. varLD: a program for quantifying variation in linkage disequilibrium patterns between populations. Bioinformatics 26, 1269–1270, https://doi.org/10.1093/bioinformatics/btq125 (2010).
20. Platt, A. et al. Genetic and regulatory mechanism of susceptibility to high-hyperdiploid acute lymphoblastic leukaemia at 10p21.2. Nat Commun 8, 1416, https://doi.org/10.1038/ncomms14166 (2017).
21. Trevino, L. R. et al. Germline genomic variants associated with childhood acute lymphoblastic leukaemia. Nat Genet 41, 1001–1005, https://doi.org/10.1038/ng.3432 (2009).
22. Papasmanuil, E. et al. Loci on 7p12.2, 10q21.2 and 14q12.1 are associated with risk of childhood acute lymphoblastic leukaemia. Nat Commun 4, 1006–1010, https://doi.org/10.1038/ncomms4300 (2013).
23. Zeng, H. et al. Associations between AT-rich interactive domain 58 gene polymorphisms and risk of childhood acute lymphoblastic leukaemia: a meta-analysis. Asian Pac J Cancer Prev 15, 6211–6217 (2014).
24. Li, H., et al. Association of the 1p36 region with risk of childhood acute lymphoblastic leukaemia in blacks and contributes to racial differences in leukemia incidence. Leukemia 31, 751–757, https://doi.org/10.1038/leu.2016.277 (2017).
25. Xu, H. et al. ARID5B genetic polymorphisms contribute to racial disparities in the incidence and treatment outcome of childhood acute lymphoblastic leukemia. J Clin Oncol 30, 751–757, https://doi.org/10.1200/JCO.2011.38.0345 (2012).
26. Walsh, K. M. et al. Associations between genome-wide Native American ancestry, known risk alleles and B-cell ALL risk in Hispanic children. Leukemia 27, 2416–2419, https://doi.org/10.1038/leu.2013.130 (2013).
27. Wang, X. et al. Interpretation of association signals and identification of causal variants from genome-wide association studies. Am J Hum Genet 86, 730–742, https://doi.org/10.1016/j.ajhg.2010.04.003 (2010).
28. Sherborne, A. L. et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. Nat Genet 42, 190–194, https://doi.org/10.1038/ng.585 (2010).
29. Vijayakrishnan, J. et al. The 9p21.3 risk of childhood acute lymphoblastic leukaemia is explained by a rare high-impact variant in CDKN2A. Sci Rep 5, 15065, https://doi.org/10.1038/srep15065 (2015).

www.nature.com/scientificreports/
36. Walsh, K. M. et al. A Heritable Missense Polymorphism in CDKN2A Confers Strong Risk of Childhood Acute Lymphoblastic Leukemia and Is Preferentially Selected during Clonal Evolution. *Cancer Res* 75, 4884–4894, https://doi.org/10.1158/0008-5472.CAN-15-1105 (2015).

37. Xu, H. et al. Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. *Nat Commun* 6, 7553, https://doi.org/10.1038/ncomms8553 (2015).

38. Chakravarti, A. Population genetics—making sense out of sequence. *Nat Genet* 21, 56–60, https://doi.org/10.1038/44882 (1999).

39. Perez-Andreu, V. et al. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. *Nat Genet* 45, 1494–1498, https://doi.org/10.1038/ng.2803 (2013).

40. Migliorini, G. et al. Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. *Blood* 122, 3298–3307, https://doi.org/10.1182/blood-2013-03-491316 (2013).

41. Tsuchida, M. et al. Long-term results of Tokyo Children's Cancer Study Group trials for childhood acute lymphoblastic leukemia, 1984–1999. *Leukemia* 24, 383–396, https://doi.org/10.1038/leu.2009.260 (2010).

42. Xu, H. et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. *J Natl Cancer Inst* 105, 733–742, https://doi.org/10.1093/jnci/djt042 (2013).

43. Walsh, K. M. et al. GATA3 risk alleles are associated with ancestral components in Hispanic children with ALL. *Blood* 122, 3385–3387, https://doi.org/10.1182/blood-2013-08-524124 (2013).

Acknowledgements

This work was supported by funding from St. Luke's Life Science Institute (Tokyo, Japan), the Children's Cancer Association of Japan, and the Japan Leukemia Research Fund. Genotyping was partially supported by Japan Society for the Promotion of Science KAKENHI Grant number 26253041. We would like to thank the patients and families participating in this study, and staff of the collaborating hospitals for their various contributions. This study made use of data from the 1000 Genomes Project (http://www.internationalgenome.org/data) and the International HapMap Project.

Author Contributions

K.Y.U., A.M., M.T., S.M., Y.I. and F.M. conceived and designed the study; Y.T., Y.A., D.H., Y.Y., T.K., Y.N., Y.T., S.O., T.I., M.Y., Kaz.K., D.T., Y.N., H.K., K.N., Ko.M., H.G., Y.S., D.M., M.K., J.T., Kat.K., M.Y., A.O. and A.M. were involved in patient recruitment and sample collection; M.T., Ke.M., T.K., F.M., T.T. and J.I. led the laboratory analyses and generation of genomic data; K.Y.U. and T.K. and conducted the statistical analysis and bioinformatics evaluations; K.Y.U. drafted the first version of the manuscript. All authors critically reviewed and edited the manuscript for intellectual content and gave final approval of the final version.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018