HHS Public Access

Author manuscript

Leukemia. Author manuscript; available in PMC 2022 September 01.

Published in final edited form as:

Leukemia. 2022 March ; 36(3): 865–868. doi:10.1038/s41375-021-01465-1.

Genome-wide trans-ethnic meta-analysis identifies novel susceptibility loci for childhood acute lymphoblastic leukemia:

Transethnic GWAS on acute lymphoblastic leukemia

Soyoung Jeon[1,2], **Adam J. de Smith**[1], **Shaobo Li**[1,2], **Minhui Chen**[1], **Tsz Fung Chan**[1], **Ivo S. Muskens**[1], **Libby M. Morimoto**[3], **Andrew T. DeWan**[4,5], **Nicholas Mancuso**[1,6,7], **Catherine Metayer**[3], **Xiaomei Ma**[5], **Joseph L. Wiemels**[1], **Charleston W.K. Chiang**[1,6]

[1.]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA

[2.]Cancer Biology and Genomics Graduate Program, Program in Biological and Biomedical Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA

[3.]Division of Epidemiology & Biostatistics, School of Public Health, University of California, Berkeley, CA

[4.]Center for Perinatal, Pediatric and Environmental Epidemiology, Yale School of Public Health, New Haven, CT

[5.]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT

[6.]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA

[7.]Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA

Genome-wide association studies (GWAS) have identified a number of risk loci for childhood acute lymphoblastic leukemia (ALL)(1–6) and estimated the heritability to be 21% (ref.(7)). The known risk loci together account for a relatively small portion of the total variance in genetic risk(7), suggesting that additional susceptibility alleles may be discovered. Furthermore, published GWASs predominantly investigated only Europeans, despite the racial and ethnic disparities in ALL incidence and outcomes(8,9).

Given this context, we performed a trans-ethnic GWAS of childhood ALL in a discovery panel of 76,317 individuals, including 3,482 cases and 72,835 controls distributed across four ethnic cohorts (African Americans, AFR; East Asian Americans, EAS; Latino Americans, LAT; Non-Latino White Americans, NLW; Supplementary Information). After

Correspondence: Joseph Leo Wiemels, Center for Genetic Epidemiology, 1450 Biggy St, Los Angeles, California, wiemels@usc.edu, phone: (323) 442-7865, fax: (323) 442-7749; Charleston W.K. Chiang, 1450 Biggy St, Los Angeles, California, charleston.chiang@med.usc.edu, phone: (323)442-8052, fax: (323) 442-7749.

Authorship

Contribution: J.L.W., C.W.K.C., A.J.D. conceived and supervised this project; S.J., S.L., M.C., T.C. performed data analysis; N.M., I.S.M., L.M.M., A.T.D., C.M., X.M. provided resources; S.J., A.J.D., C.W.K.C., J.L.W. wrote the manuscript with input from all the coauthors. Conflict-of-interest disclosure: The authors declare no competing interests.

Competing Interests: The authors declare no competing interests.

quality control filtering, our dataset consisted of 124, 318, 1,878, 1,162 cases and 2,067, 5,017, 8,410, 57,341 controls in AFR, EAS, LAT and NLW, respectively (Table S1; Figure S1, Supplementary Information). Furthermore, we tested the association at 7,628,894 imputed SNPs, including low frequency (minor allele frequency, MAF, between 1-5%) variants that were not previously systematically tested. We aggregated summary statistics across the four ethnic groups in a fixed-effect meta-analysis (Figure S1). The genomic control inflation factor was 1.022 after excluding 16 previously reported ALL-associated loci (Table S2), suggesting our meta-analysis was reasonably robust to any confounding due to population stratification (Figure S2).

Of 16 previously published risk loci, all were nominally associated with ALL ($P<0.05$) or have a SNP nearby with strong association (Table S2). Most of these risk SNPs showed consistent direction of effects across ethnic groups and little evidence of heterogeneity (but note *C5orf56* and *TLE1* in Table S3). Given the larger sample size and trans-ethnic analysis, the best associated variants in our analysis may reflect the more likely causal/shared association across populations. More importantly, we discovered three putatively novel susceptibility loci: one at 6q23 and two at 10q21 (Table S4, Figure S2). The strongest association signal in 6q23 is at rs9376090 ($P= 8.23 \times 10^{-9}$, OR=1.27) in the intergenic region between *MYB* and *HBS1L* (Figure 1A). A locus in 10q21 was identified with the lead SNP rs9415680 ($P=7.27 \times 10^{-8}$, OR=1.20), within a broad association peak and apparently long-range LD with SNPs overlapping *NRBF2, JMJD1C*, and *REEP3* (Figure 1B). A second 10q21 locus was identified 5Mb upstream with lead SNP rs10998283 ($P=3.92\times10^{-8}$, OR=1.15) in an intronic region in *TET1* (Figure 1C). Each of the three putatively novel loci harbors genes and/or variants with a role in hematopoiesis and leukemogenesis, and is within larger chromatin regions containing several genes in a B-lymphoblastoid cell line that mirrors the differentiation state of the majority of childhood ALL (Supplemental Information, Figures S3–S5). We found little difference in association with different ALL subtypes (Table S5).

We tested for association of the three novel variants and their LD proxies (with $P<5\times10^{-7}$; n=141) in the independent COG/WTCCC and CCLS replication cohorts (Figure S1, Supplementary Information). For *MYB/HBS1L,* where the association with ALL was driven by NLW in the discovery cohort, we replicated the signal in COG/WTCCC (rs9376090, $P_{COG}=4.87\times10^{-3}$, $P_{COG+discovery\ analysis}=1.23\times10^{-10}$; Table S6), but not in CCLS likely owing to its small NLW cohort. For *TET1*, in which the association was driven by LAT in the discovery cohort, three of the four SNPs with $P<5\times10^{-7}$ in the discovery cohort nominally replicated in CCLS (lead SNP rs7922602; $P_{CCLS}=3.04\times10^{-2}$, $P_{CCLS+discovery}=6.81\times10^{-9}$; Table S6). The *TET1* SNPs did not replicate in COG/WTCCC, suggesting we may have over-estimated its effect in NLW in the discovery cohort. SNPs in the *NRBF2/JMJD1C* locus did not replicate in either replication cohort.

Conditional analyses adjusting for the lead SNP at each locus identified a secondary signal in four out of the 16 previously known loci (Table S7, Figure S6). In all cases, the LD between the secondary association and the top association in the locus are low (Table S7). Additional secondary associations in *CDKN2A* and *IZKF1* loci were previously noted(7). In *CEBPE* (rs60820638, $P=5.38\times10^{-8}$) and 17q12 (rs12944882, $P=7.71\times10^{-10}$),

these secondary signals represent novel associations. In particular, at *CEBPE*, previous reports suggested multiple correlated variants with functional evidence(10,11) . Our analysis is consistent with the two previous variants (rs2239635 and rs2239630) being or tagging the same underlying signal, while the new association we identified (rs60820638) is independent (Table S8).

To assess the combined effect of all identified risk alleles for ALL, we trained a PRS model based on 18 previously known ALL-associated SNPs or 23 known and novel SNPs in our discovery cohort , CCRLP (Supplementary Information). We tested the PRS models in the NLW and LAT individuals, the largest subcohorts, in the independent CCLS and COG/WTCCC cohorts (Figure S1). The PRS models were significantly associated with case-control status in all groups, and the predictive accuracy as measured by AUC are similar between NLW and LAT, at around 67-68% (Table S9), consistent with the expectation that trans-ethnic meta-analysis will enable PRS to be more transferrable between populations. We found that the shape of the PRS distribution appears similar between LAT and NLW individuals from CCRLP, but the scores in LAT are significantly shifted to the right compared to the scores in NLW (mean of 5.101 and 4.641 respectively, Welch t-test P = $1.3x10^{-122}$, Figure 2A). The separation between LAT and NLW was also observed when scores were stratified by case-control status (P=$3.956\ x10^{-58}$ and $1.493\ x10^{-78}$ among cases and controls, respectively; Figure 2C), and was replicated in CCLS (P=$4.596\ x10^{-51}$; Figure 2B). Results from our PRS analyses support that differences in allele frequency of ALL risk loci between populations may partly explain the increased ALL risk in LAT relative to NLW children.

In CCLS, where effect size estimates are expected to be less biased by winner's curse, we found that known risk variants collectively accounted for similar proportions of familial relative risk in LAT and NLW (~23-24%; Table S10). The heritability estimates of ALL attributable to common imputed SNPs (MAF ≥ 0.05), however, differed between LAT and NLW. Using either the GCTA-LDMS framework(12) or the phenotype-correlation-genotype-correlation (PCGC) regression framework(13), heritability of ALL was consistently estimated to be ~20% in NLW (similar to previous estimates (7)), but ranged from 4% to 11% in LAT (Table S11). The difference in heritability estimates contrasted strongly against the observation of similar estimated effect sizes per SNP (r2 = 0.819; Figure S7) and similar familial relative risks explained (Table S10) among GWAS loci between LAT and NLW, suggesting that there may be model instability or misspecification due to the admixed nature of LAT when estimating heritability(14) or there may be differences in environmental exposures. Despite this, we estimated the genetic correlation of ALL between NLW and LAT to be high (rG=0.714 ± standard error 0.130) but significantly different from 1 (P=0.014, Table S12), and that approximately 32.5% of SNPs inferred to be causal are shared between NLW and LAT (Supplementary Information).

Because low frequency variants (MAF between 1-5%) are expected to be well-imputed in NLW, we estimate that the inclusion of low frequency variants increased the estimated heritability in NLW to 29.8 ± 4.3% (~16.2% due to common variants, 13.5% due to low frequency variants; Table S13). Taking advantage of the admixed nature of LAT, whereby ancestry segments could capture effects beyond that directly attributable to assayed SNPs

(such as the estimate from GCTA-LDMS), we also adopted an approach (15) to estimate the total narrow-sense heritability for ALL in LAT to be 37.3 ± 6.9% (Supplementary Information). Taken together, multiple lines of evidence suggest that increasing sample sizes will identify additional low frequency associations to ALL in the future.

In summary, we performed the largest trans-ethnic meta-analysis GWAS of childhood ALL to date, identifying three putatively novel susceptibility loci (although we could not replicate the association at *NRBF2* | *JMJD1C* locus) and two additional independent risk associations at previously reported loci. Our analysis suggests that the known and novel ALL risk alleles together explained ~24% of familial relative risk in both NLW and LAT populations, and that the trans-ethnic PRS we constructed, although relatively simple and utilizing only the genome-wide associated variants, performed similarly in both NLW and LAT in predicting ALL (AUC ~67-68%). Our results also suggest multiple avenues for future studies of ALL. First, use of larger study cohorts will permit identification of additional alleles at lower frequency, particularly given the significant proportion of heritability explained by low frequency variants in NLW. Increasing sample sizes from AFR and EAS cohorts will improve our understanding of the genetic architecture of ALL in these populations, including heritability estimates and efficacy of PRS models. Second, ethnic-specific studies for ALL are urgent for discovery of ancestry-specific associations that may be missed in a trans-ethnic GWAS. Indeed, several the previously known or putatively novel loci reported here show significant heterogeneity across four ethnic groups (Table S3, S4). Finally, our discovery cohort has limited subtype information. Future studies should focus on disentangling the different subtypes of ALL, and to study other aspects of the disease pathogenesis such as disease progression or risk of relapse.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This work was supported by research grants from the National Institutes of Health (R01CA155461, R01CA175737, R01ES009137, P42ES004705, P01ES018172, P42ES0470518, R24ES028524 and R35GM142783) and the Environmental Protection Agency (RD83451101), United States. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and the EPA. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C awarded to the Cancer Prevention Institute of California, contract HHSN261201000035C awarded to the University of Southern California, and contract HHSN261201000034C awarded to the Public Health Institute; and the Centers for Disease Control and Prevention's National Program of Cancer Registries, under agreement U58DP003862-01 awarded to the California Department of Public Health. The biospecimens and/or data used in this study were obtained from the California Biobank Program, (SIS request #26), Section 6555(b), 17 CCR. The California Department of Public Health is not responsible for the results or conclusions drawn by the authors of this publication. We thank Hong Quach and Diana Quach for DNA isolation support. We thank Martin Kharrazi, Robin Cooley, and Steve Graham of the California Department of Public Health for advice and logistical support. We thank Eunice Wan, Simon Wong, and Pui Yan Kwok at the UCSF Institute of Human Genetics Core for genotyping support. This study makes use of data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475. Genotype data for COG ALL cases are available for download from dbGaP (Study Accession: phs000638.v1.p1).

Data came from a grant, the Resource for Genetic Epidemiology Research in Adult Health and Aging (RC2 AG033067; Schaefer and Risch, PIs) awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics. The RPGEH was supported by grants from the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente Northern California, and the Kaiser Permanente National and Northern California Community Benefit Programs. The RPGEH and the Resource for Genetic Epidemiology Research in Adult Health and Aging are described here: https://divisionofresearch.kaiserpermanente.org/genetics/rpgeh/rpgehhome. For recruitment of subjects enrolled in the CCLS replication set, the authors gratefully acknowledge the clinical investigators at the following collaborating hospitals: University of California Davis Medical Center (Dr. Jonathan Ducore), University of California San Francisco (Drs. Mignon Loh and Katherine Matthay), Children's Hospital of Central California (Dr. Vonda Crouse), Lucile Packard Children's Hospital (Dr. Gary Dahl), Children's Hospital Oakland (Dr. James Feusner), Kaiser Permanente Roseville (formerly Sacramento) (Drs. Kent Jolly and Vincent Kiley), Kaiser Permanente Santa Clara (Drs. Carolyn Russo, Alan Wong, and Denah Taggart), Kaiser Permanente San Francisco (Dr. Kenneth Leung), and Kaiser Permanente Oakland (Drs. Daniel Kronish and Stacy Month). The authors additionally thank the families for their participation in the California Childhood Leukemia Study (formerly known as the Northern California Childhood Leukemia Study). Finally, the authors acknowledge the Center for Advanced Research Computing (CARC; https://carc.usc.edu) at the University of Southern California for providing computing resources that have contributed to the research results reported within this publication.

REFERENCES

1. Vijayakrishnan J, Kumar R, Henrion MYR, Moorman AV, Rachakonda PS, Hosen I, et al. A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. Leukemia. 2017 Mar;31(3):573–9. [PubMed: 27694927]
2. Wiemels JL, Walsh KM, de Smith AJ, Metayer C, Gonseth S, Hansen HM, et al. GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. Nat Commun. 2018 Dec;9(1):286. [PubMed: 29348612]
3. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. Nat Genet. 2009 Sep;41(9):1006–10. [PubMed: 19684604]
4. Perez-Andreu V, Roberts KG, Harvey RC, Yang W, Cheng C, Pei D, et al. Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. Nat Genet. 2013 Dec;45(12):1494–8. [PubMed: 24141364]
5. Treviño LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. Nat Genet. 2009 Sep;41(9):1001–5. [PubMed: 19684603]
6. Xu H, Yang W, Perez-Andreu V, Devidas M, Fan Y, Cheng C, et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. J Natl Cancer Inst. 2013 May 15;105(10):733–42. [PubMed: 23512250]
7. Vijayakrishnan J, Qian M, Studd JB, Yang W, Kinnersley B, Law PJ, et al. Identification of four novel associations for B-cell acute lymphoblastic leukaemia risk. Nat Commun. 2019 Dec;10(1):5348. [PubMed: 31767839]
8. Giddings BM, Whitehead TP, Metayer C, Miller MD. Childhood leukemia incidence in California: High and rising in the Hispanic population: Hispanic Childhood Leukemia Incidence. Cancer. 2016 Sep 15;122(18):2867–75. [PubMed: 27351365]
9. Lim JY-S, Bhatia S, Robison LL, Yang JJ. Genomics of racial and ethnic disparities in childhood acute lymphoblastic leukemia. Cancer. 2014 Apr 1;120(7):955–62. [PubMed: 24382716]
10. Wiemels JL, de Smith AJ, Xiao J, Lee S-T, Muench MO, Fomin ME, et al. A functional polymorphism in the CEBPE gene promoter influences acute lymphoblastic leukemia risk through interaction with the hematopoietic transcription factor Ikaros. Leukemia. 2016 May;30(5):1194–7. [PubMed: 26437776]
11. Studd JB, Yang M, Li Z, Vijayakrishnan J, Lu Y, Yeoh AE-J, et al. Genetic predisposition to B-cell acute lymphoblastic leukemia at 14q11.2 is mediated by a CEBPE promoter polymorphism. Leukemia. 2019 Jan;33(1):1–14. [PubMed: 29977016]
12. The LifeLines Cohort Study, Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat Genet. 2015 Oct;47(10):1114–20. [PubMed: 26323059]

13. Golan D, Lander ES, Rosset S. Measuring missing heritability: Inferring the contribution of common variants. Proc Natl Acad Sci. 2014 Dec 9;111(49):E5272–81. [PubMed: 25422463]

14. Steinsaltz D, Dahl A, Wachter KW. On Negative Heritability and Negative Estimates of Heritability. Genetics. 2020 Jun;215(2):343–57. [PubMed: 32291292]

15. Zaitlen N, Pasaniuc B, Sankararaman S, Bhatia G, Zhang J, Gusev A, et al. Leveraging population admixture to characterize the heritability of complex traits. Nat Genet. 2014 Dec;46(12):1356–62. [PubMed: 25383972]

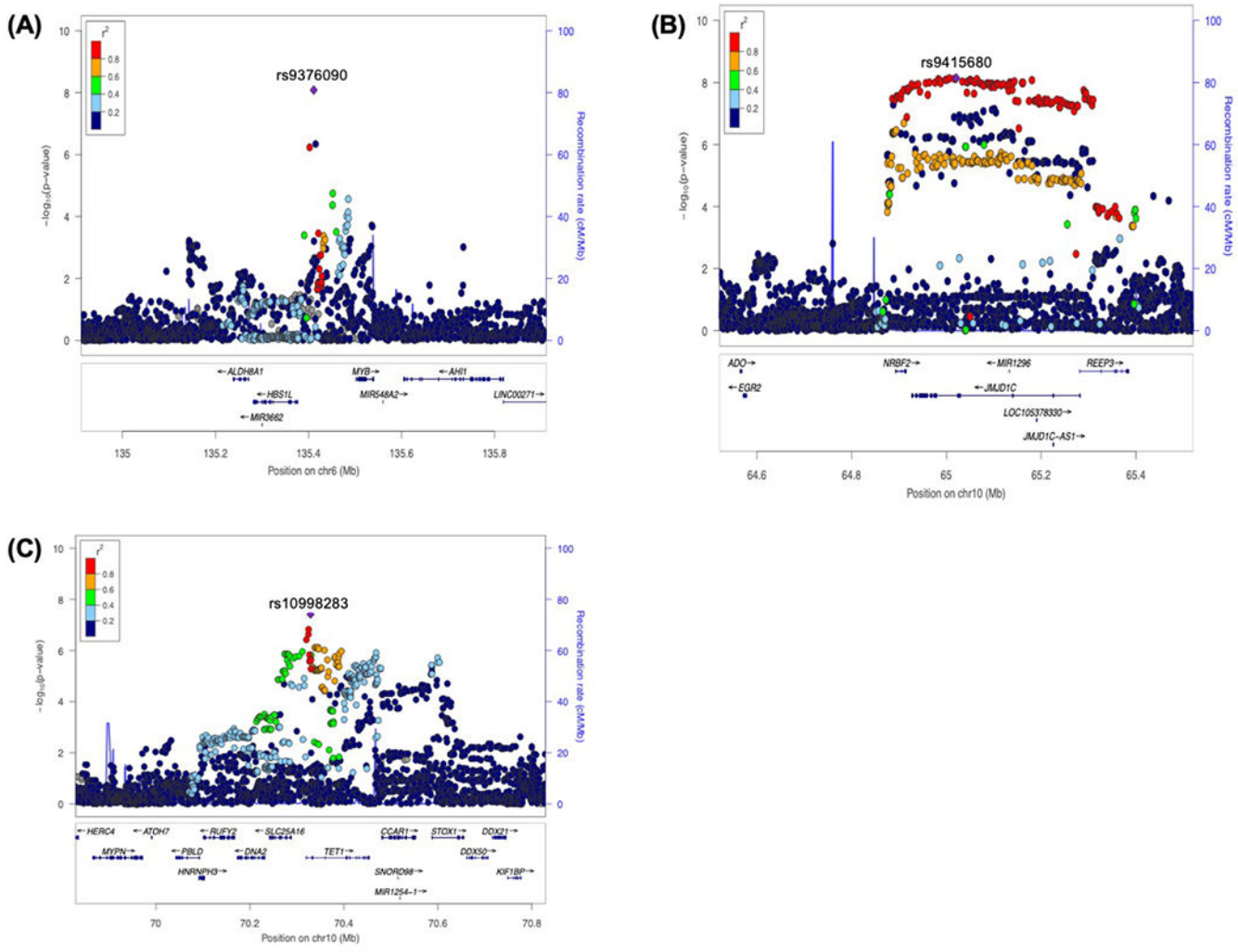

Figure 1. Novel loci associated with childhood ALL in trans-ethnic meta-analysis.

LocusZoom plots showing 1 Mb region around the identified loci near (A) MYB/HBS1L on chr6, (B) NRBF2/JMJD1C on chr10, and (C) TET1 on chr10 are shown. Diamond symbol indicates the lead SNP in each locus. Color of remaining SNPs is based on linkage disequilibrium (LD) as measured by r2 with the lead SNP in non-Latino white. All coordinates in x-axis are in hg19.

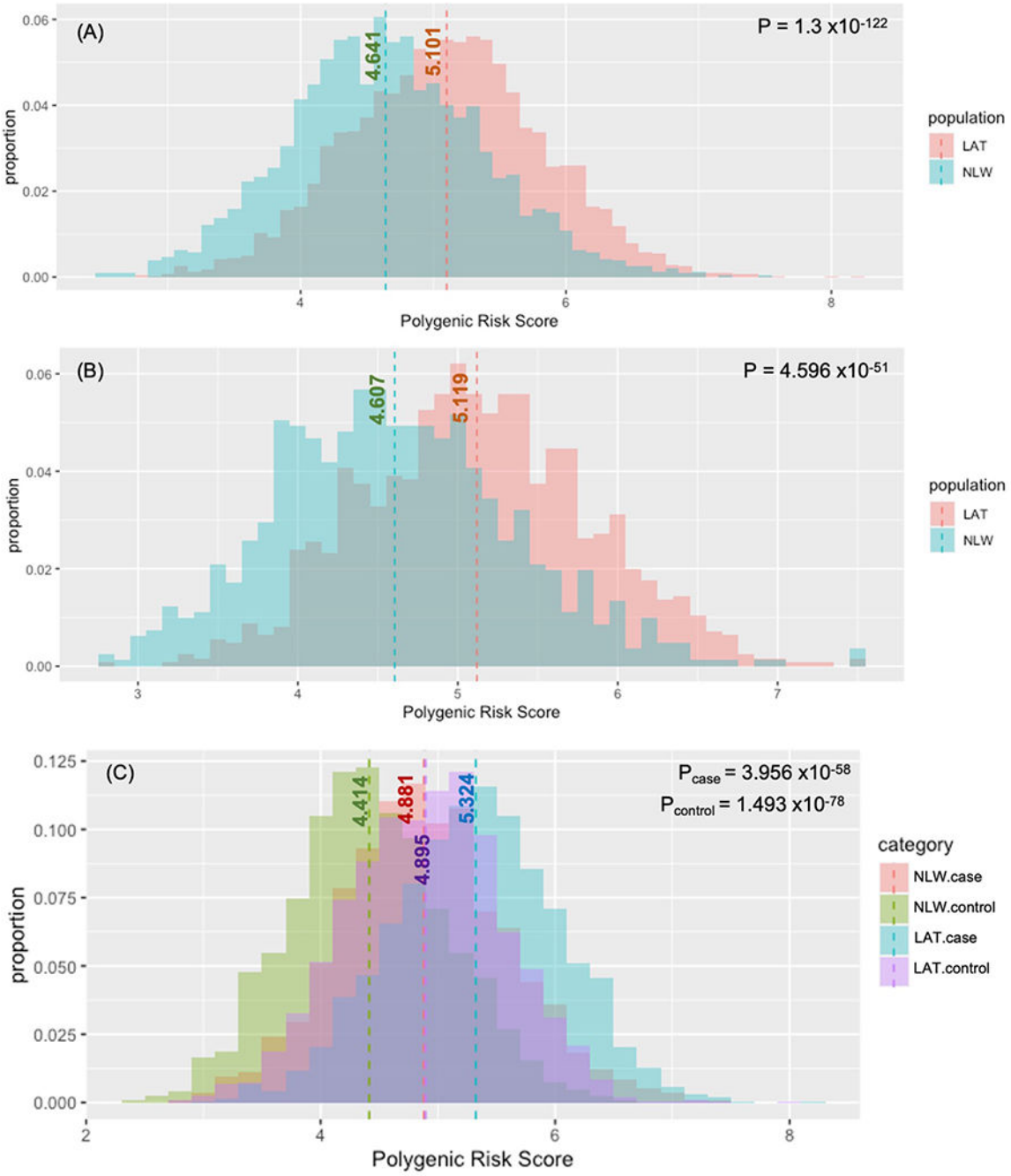

Figure 2. Polygenic Risk Score (PRS) distribution based on GWAS loci for ALL.

We compared the PRS distribution between LAT and NLW cohorts in (A) CCRLP and (B) CCLS cohorts. The distributions of PRS are consistent with a normal distribution (Kolmogorov[-Smirnov P = 0.918 and 0.303 for LAT and NLW, respectively) and the shape between LAT and NLW are similar (standard deviation of 0.728 and 0.735 respectively; F-Test P = 0.633). In (C) We further stratified the PRS in CCRLP cohort by case/control status. The population mean is indicated with vertical dash lines with the mean score shown.

P-values on the right upper corner of each graph is from one-sided t-test comparing the difference in PRS between LAT and NLW overall or within cases and controls.