ARTICLE - Acute Lymphoblastic Leukemia

Evaluation of the genetic basis of familial-associated early-onset hematologic cancers in an ancestral/ethnically diverse population

Qianxi Feng,[1] Keren Xu,[1] Mancy Shah,[2] Shaobo Li,[1] Andrew D. Leavitt,[3] Lucy A. Godley,[2] Adam J. de Smith[1] and Joseph L. Wiemels[1]

[1]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA; [2]Division of Hematology/Oncology, Department of Medicine, and the Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL and [3]Departments of Medicine and Laboratory Medicine, University of California, San Francisco, CA, USA

Correspondence: J.L. Wiemels
wiemels@usc.edu

Received: September 20, 2023.
Accepted: December 29, 2023.
Early view: January 11, 2024.

https://doi.org/10.3324/haematol.2023.284224

Abstract

Genetic predisposition to hematologic malignancies has historically been addressed utilizing patients recruited from clinical trials and pedigrees constructed at major treatment centers. Such efforts leave unexplored the genetic basis of variations in risk by race/ethnic group shown in population-based surveillance data where cancer registration, compulsory by law, delivers universal enrollment. To address this, we performed exome sequencing on DNA isolated from newborn bloodspots derived from sibling pairs with early-onset cancers across California in which at least one of the siblings developed a hematologic cancer, using unbiased recruitment from the full state population. We identified pathogenic/likely pathogenic (P/LP) variants among 1,172 selected cancer genes that were private or present at low allele frequencies in reference populations. Within 64 subjects from 32 families, we found 9 LP variants shared between siblings, and an additional 7 such variants in singleton children (not shared with their sibling). In 8 of the shared cases, the ancestral origin of the local haplotype that carries P/LP variants matched the dominant global ancestry of study participant families. This was the case for Latino sibling pairs on *FLG* and *CBLB*, non-Latino White sibling pairs in *TP53* and *NOD2*, and a shared *GATA2* variant for a non-Latino Black sibling pair. A new inherited mutation in *HABP2* was identified in a sibling pair, one with diffuse large B-cell lymphoma and the other with neuroblastoma. Overall, the profile of P/LP germline variants across ancestral/ethnic groups suggests that rare alleles contributing to hematologic diseases originate within their race/ethnic origin parental populations, demonstrating the value of this discovery process in diverse, population-based registries.

Online Supplementary Tables included with “Evaluation of the genetic basis of familial-associated early- onset hematologic cancers in an ancestral/ethnically diverse population” by Feng, Q., et al.

Online Supplementary Table 1. List of genes evaluated for pathogenicity.

Online Supplementary Table 2. Variants of uncertain significance (VUS) shared between family members.

Online Supplementary Table 3. Pathogenic/Likely pathogenic (P/LP) variants shared between family members: Results using manual curation after software variant calling.

Online Supplementary Table 4. Pathogenic/Likely pathogenic (P/LP) variants and variants of uncertain significance (VUS) shared between family members: Initial predictions based on Varsome and Clinvar (via PeCanPIE), and Manual Curation updated variant designations. Supp Table 2b includes the manual variant curation decision process.

Online Supplementary Table 5. Pathogenic/Likely pathogenic (P/LP) variants not shared between family members.

Online Supplementary Table 6. Basic clinical diagnostic information of all families included in this study.

Supplemental Methods

DNA preparation and sequencing

DNA used for augmented whole exome sequencing (WES)[1] was isolated from neonatal dried blood samples obtained from the California Biobank Program[2] using Beckman GenFind v3 reagents on an Eppendorf robotic sample handling platform. Uniquely barcoded samples underwent WES on the IDT xGen Exome V1 plus spike-in of a small panel of clinically relevant probes that cover additional non-coding loci where predisposition alleles reside (detailed in[1]). Approximately 250 million paired end reads, each 100 bp in length, were generated for each sample.

Mapping and variant identification

The Genome Analysis ToolKit (GATK) pipeline for germline short variant (SNVs + indels) discovery was used for mapping and variant calling[3-5], based on the GRCh37 assembly. Resulting gene sequence variations stored in variant call format (VCF) files were annotated with ANNOVAR[6]. Variants with alternative allele reading depth ≤5, or variant allele fraction ≤0.2 were excluded. Variants with quality by depth>2 and genotype quality>10 were included.

We filtered all variants for minor allele frequency (MAF) in reference populations in the Genome Aggregation Database (gnomAD)[7, 8] and the 1000 Genomes Project (1KG)[9]. Rare variants with both global and population-specific allele frequency<=0.001 in the exome sequencing data in gnomAD and 1KG were included in the filtered VCF file.

Identification of Down syndrome (DS)

The presence of DS (trisomy 21) for each subject was identified by comparing the sequencing read ratio for chromosome 21 to all other chromosomes. Then, any subject with number of reads for chromosome 21: other ratio greater than the mean + 2*standard deviation (SD) of the reads of all other subjects on this chromosome were identified to have constitutive trisomy 21.

Annotation of pathogenicity

We used the Pediatric Cancer Variant Pathogenicity Information Exchange (PeCanPIE) Medal Ceremony pipeline for the initial identification of P/LP variants across all sequenced siblings. PeCanPIE works by first sifting through variants in sequencing data, and then annotating the pathogenicity of the variants based on American College of Medical Genetics and Genomics (ACMG)[10]/Association for Molecular Pathology (AMP) guidelines[11]. The potential pathogenicity of the variants is classified into three tiers (gold, silver, and bronze)[12]. We examined variants on a list of 1173 genes that are reported to be cancer/immunodeficiency/nonmalignant hematological-related genes (n=986)[12, 13], pediatric cancer predisposition genes (n=162)[14, 15], tumor suppressor genes, tyrosine kinase genes, or cancer genes classified based on their recurrent somatic mutation in cancer (n=565)[16], and/or Hodgkin lymphoma-related genes identified from genome-wide association or sequencing studies (n=327)[17-25] (Online Supplementary Table 1).

Then, the pathogenicity of 'gold' and 'silver' medal variants identified by PeCanPIE was cross-checked with VarSome[26], a search engine for variants in the human genome that classifies different pathogenicity categories according to ACMG/AMP guidelines by incorporating information from external databases and risk prediction scores from multiple *in silico* algorithms. The pathogenicity of each variant is annotated as 'pathogenic', 'likely pathogenic', 'likely benign', 'benign' or 'uncertain significance' according to ACMG/AMP guidelines.

Variants that have a PeCanPIE 'gold' medal, or VarSome annotation as 'pathogenic' or 'likely pathogenic' are referred to as 'P/LP' in the subsequent analyses. Other PeCanPIE 'silver' medal variants are referred to as variants of unknown significance (VUS) in the subsequent analyses. All P/LP/VUS variants were then manually annotated manually for pathogenicity by examining multiple sources including literature reports, defined mutational hotspots, database reports, and functional studies. Initial "in silico" variant classifications (by PeCanPIE) are shown in Supplementary Table 2 only; all manually curated alleles are displayed in other key tables (Table 2, Figure 1, Online Supplementary Tables 2, 3, 5).

All putative P/LP variants and VUS that were shared by both siblings were inspected visually with the Integrative Genomics Viewer[27] to ensure adequate sequencing depth, and the percentage of alternative allele reads was recorded for each variant (and displayed in tables). We also performed manual inspection of putative P/LP variants that were found in only one sibling and subjected these variants to the same manual curation process.

Ancestry of variants

To evaluate if the variants originated from a specific genetic ancestry, we examined the global ancestries of each study subject further, and the local ancestries surrounding each variant. The ancestries were classified into 5 superpopulations: European (EUR), African (AFR), Amerindian (AMR), East Asian (EAS), and South Asian (SAS). RFMix[28] was used to determine the global and local ancestries.

We used gene sequence variations from the 1KG[9] and Human Genome Diversity Project (HGDP)[29] to construct a reference panel for the ancestry analysis with RFMix. First, ADMIXTURE[30] was used to identify the 1KG and HGDP subjects with a 'pure' global ancestry (100%EUR/100%AFR/100%AMR/100%EAS/100%SAS). A total of 345 ancestrally unmixed subjects (69 AFR, 69 AMR, 69 EAS, 69 EUR, 69 SAS – with AMR being the group with the fewest available ancestrally uniform subjects thus limiting the size of each reference group) were included in the reference panel for RFMix. We then mapped the local gene sequence variations of all study participants to the gene sequence variations of these reference subjects to determine the global ancestry of each chromosome and the local ancestry of each variant. The global ancestry of each subject was calculated by averaging the global ancestry of each chromosome for that subject. The local ancestry of a common SNP that is closest to the variant of interest was deemed to be the local ancestry of that variant. A full listing of all family clinical information is presented in Online Supplementary Table 6.

Acknowledgements

This study used biospecimens from the California Biobank Program (SIS #1102). Research was approved by the California State Committee for Protection of Human Subjects and the Vital Statistics Advisory Committee. We are respectfully unable to share raw, individual-level genomic and genome-wide DNA methylation data reported in this study, which are property of the State of California. Should we be contacted regarding individual-level data contributing to the findings reported in this study, inquiries will be directed to the California Department of Public Health Institutional Review Board to establish an approved protocol to

utilize the data, which cannot otherwise be shared peer-to-peer. The State of California has provided guidance on data sharing per the following statement: “California has determined that researchers requesting the use of California Biobank biospecimens for their studies will need to seek an exemption from NIH or other granting or funder requirements regarding the uploading of study results into an external bank or repository (including into the NIH dbGaP or other bank or repository). This applies to any uploading of genomic data and/or sharing of these biospecimens or individual data derived from these biospecimens. Such activities have been determined to violate the statutory scheme of California Health and Safety Code Section 124980 (j), 124991 (b), (g), (h), and 103850 (a) and (d), which protect the confidential nature of biospecimens and individual data derived from biospecimens. All investigators seeking to use California specimens for projects or grant-related activities that require or seek such sharing (at the NIH or elsewhere) must seek an exemption from genomic data sharing requirements. If such an exemption is not secured, samples will not be released to an investigator. Investigators may agree to share aggregate data on SNP frequency and their associated P-values with other investigators and may upload such frequencies into repositories including the NIH dbGaP repository providing (a) the denominator from which the data is derived includes no fewer than 20,000 individuals; (b) no cell count is for <5 individuals; and (c) no correlations or linkage probabilities between SNPs are provided.”

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 ideas and opinions expressed herein are those of the author(s), and no endorsement by the California Department of Public Health, the National Cancer Institute, or the Centers for Disease Control and Prevention or their contractors and subcontractors is intended or should be inferred.

1. Feurstein S, Trottier AM, Estrada-Merly N, Pozsgai M, McNeely K, Drazer MW, Ruhle B, Sadera K, Koppayi AL, Scott BL, Oran B, Nishihori T, et al. Germ line predisposition variants occur in myelodysplastic syndrome patients of all ages. Blood 2022;140: 2533-48.
2. California Biobank Program, vol. 2021: California Depertment of Public Health, 2018.
3. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20: 1297-303.
4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43: 491-8.
5. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43: 11 0 1- 0 33.
6. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38: e164.
7. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, et al. Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2021;590: E53.
8. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581: 434-43.
9. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature 2015;526: 68-74.
10. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE, Molecular Subcommittee of the ALQAC. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med 2008;10: 294-300.
11. Lyon E, Temple-Smolkin RL, Hegde M, Gastier-Foster JM, Palomaki GE, Richards CS. An Educational Assessment of Evidence Used for Variant Classification: A Report of the Association for Molecular Pathology. J Mol Diagn 2022.
12. Edmonson MN, Patel AN, Hedges DJ, Wang Z, Rampersaud E, Kesserwan CA, Zhou X, Liu Y, Newman S, Rusch MC, McLeod CL, Wilkinson MR, et al. Pediatric Cancer

Variant Pathogenicity Information Exchange (PeCanPIE): a cloud-based platform for curating and classifying germline variants. Genome Res 2019;29: 1555-65.

13. Kraft IL, Godley LA. Identifying potential germline variants from sequencing hematopoietic malignancies. Blood 2020;136: 2498-506.

14. Grobner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, Johann PD, Balasubramanian GP, Segura-Wang M, Brabetz S, Bender S, Hutter B, et al. The landscape of genomic alterations across childhood cancers. Nature 2018;555: 321-7.

15. Grobner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, Johann PD, Balasubramanian GP, Segura-Wang M, Brabetz S, Bender S, Hutter B, et al. Author Correction: The landscape of genomic alterations across childhood cancers. Nature 2018;559: E10.

16. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, Wilkinson MR, Vadodaria B, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med 2015;373: 2336-46.

17. Cozen W, Li D, Best T, Van Den Berg DJ, Gourraud PA, Cortessis VK, Skol AD, Mack TM, Glaser SL, Weiss LM, Nathwani BN, Bhatia S, et al. A genome-wide meta-analysis of nodular sclerosing Hodgkin lymphoma identifies risk loci at 6p21.32. Blood 2012;119: 469-75.

18. Cozen W, Timofeeva MN, Li D, Diepstra A, Hazelett D, Delahaye-Sourdeix M, Edlund CK, Franke L, Rostgaard K, Van Den Berg DJ, Cortessis VK, Smedby KE, et al. A meta-analysis of Hodgkin lymphoma reveals 19p13.3 TCF3 as a novel susceptibility locus. Nat Commun 2014;5: 3856.

19. Best T, Li D, Skol AD, Kirchhoff T, Jackson SA, Yasui Y, Bhatia S, Strong LC, Domchek SM, Nathanson KL, Olopade OI, Huang RS, et al. Variants at 6q21 implicate PRDM1 in the etiology of therapy-induced second malignancies after Hodgkin's lymphoma. Nat Med 2011;17: 941-3.

20. Enciso-Mora V, Broderick P, Ma Y, Jarrett RF, Hjalgrim H, Hemminki K, van den Berg A, Olver B, Lloyd A, Dobbins SE, Lightfoot T, van Leeuwen FE, et al. A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). Nat Genet 2010;42: 1126-30.

21. Frampton M, da Silva Filho MI, Broderick P, Thomsen H, Forsti A, Vijayakrishnan J, Cooke R, Enciso-Mora V, Hoffmann P, Nothen MM, Lloyd A, Holroyd A, et al. Variation at 3p24.1 and 6q23.3 influences the risk of Hodgkin's lymphoma. Nat Commun 2013;4: 2549.

22. Law PJ, Sud A, Mitchell JS, Henrion M, Orlando G, Lenive O, Broderick P, Speedy HE, Johnson DC, Kaiser M, Weinhold N, Cooke R, et al. Genome-wide association analysis of chronic lymphocytic leukaemia, Hodgkin lymphoma and multiple myeloma identifies pleiotropic risk loci. Sci Rep 2017;7: 41071.

23. Rotunno M, McMaster ML, Boland J, Bass S, Zhang X, Burdett L, Hicks B, Ravichandran S, Luke BT, Yeager M, Fontaine L, Hyland PL, et al. Whole exome sequencing in families at high risk for Hodgkin lymphoma: identification of a predisposing mutation in the KDR gene. Haematologica 2016;101: 853-60.

24. Sud A, Kinnersley B, Houlston RS. Genome-wide association studies of cancer: current insights and future perspectives. Nat Rev Cancer 2017;17: 692-704.

25. Urayama KY, Jarrett RF, Hjalgrim H, Diepstra A, Kamatani Y, Chabrier A, Gaborieau V, Boland A, Nieters A, Becker N, Foretova L, Benavente Y, et al. Genome-wide association study of classical Hodgkin lymphoma and Epstein-Barr virus status-defined subgroups. J Natl Cancer Inst 2012;104: 240-53.

26. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, Massouras A. VarSome: the human genomic variant search engine. Bioinformatics 2019;35: 1978-80.

27. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol 2011;29: 24-6.

28. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet 2013;93: 278-88.

29. Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. Nat Rev Genet 2005;6: 333-40.

30. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19: 1655-64.