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

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

Received: September 20, 2023.
Accepted: December 29, 2023.


Publisher’s Disclaimer.
E-publishing ahead of print is increasingly important for the rapid dissemination of science. Haematologica is, therefore, E-publishing PDF files of an early version of manuscripts that have completed a regular peer review and have been accepted for publication. E-publishing of this PDF file has been approved by the authors. After having E-published Ahead of Print, manuscripts will then undergo technical and English editing, typesetting, proof correction and be presented for the authors’ final approval; the final version of the manuscript will then appear in a regular issue of the journal. All legal disclaimers that apply to the journal also pertain to this production process.
Evaluation of the genetic basis of familial-associated early-onset hematologic cancers in an ancestral/ethnically diverse population

Qianxi Feng¹, Keren Xu¹, Mancy Shah², Shaobo Li¹, Andrew D. Leavitt³, Lucy A. Godley², Adam J. de Smith¹, and Joseph L. Wiemels¹

Affiliations of authors:
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
³ Departments of Medicine and Laboratory Medicine, University of California, San Francisco, San Francisco, CA

Acknowledgements

Funded by the V Foundation for Cancer Research (FP067172, ADL, LAG, JLW). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

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.

Disclosure of Conflicts of Interest

All authors declare no competing interests.

Authorship Contributions

All authors declare no competing interests. QF and KX performed primary data analysis with help from SL. MS performed manual variant classifications. AL, LAG, and JLW conceived the study and acquired funding. AJdS and JLW supervised statistical analysis, with help from MS and LAG. QF wrote the first draft, and JLW, LAG, and AJdS completed the writing team. All authors read and approved the final draft.

Data sharing statement: Sharing of raw and processed whole exome sequencing data from California Biobank specimens is restricted by California state statutory laws and IRB (see note in Acknowledgements). Results from specific genes or regions may be requested from wiemels@usc.edu.

Running Title: Familial Germline Pathogenic Alleles and Hematologic Malignancies

Correspondence to:
Joseph L. Wiemels
Professor, Department of Preventive Medicine
Center for Genetic Epidemiology
1450 Biggy Street, NRT-1506
Los Angeles, CA 90089
(323) 442-7865
wiemels@usc.edu

Key Points

Ancestral-origin pathogenic alleles contribute to familial predisposition in hematologic malignancies in a population-based family study.
Population-based investigations can reveal the genetic and environmental basis of hematologic cancers in multiethnic populations.

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 1172 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 eight 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.
Introduction

Hematologic malignancies (including leukemias, lymphomas, and multiple myeloma) are the most commonly diagnosed early-onset cancers among children, adolescents, and young adults\(^1\). Inherited and de novo mutation of genes within critical cell development and growth/signaling pathways are central oncogenic events in the pathogenesis of hematologic cancers\(^2\). Pathogenic/likely pathogenic (P/LP) germline variants were identified in approximately 10% of pediatric hematologic cancer patients regardless of family history\(^3-6\). The risk of early-onset cancers (of any type) diagnosed under 26 years of age is 2.97 times higher among siblings and mothers who have a proband with hematologic cancer in the same family\(^7\) indicating that inherited germline variants may contribute to this excessive cancer risk. This risk varies among ancestral/ethnic groups\(^7\) – suggesting that predisposition variants may vary by identity or frequency among groups.

Among early-onset hematologic malignancies, acute lymphoblastic leukemia is the most common, accounting for approximately 25% of all cancers diagnosed in children; and other lymphoid malignancies account for an additional 10% of all cancers\(^1,\, 8\). In our examination of linked population registries in California, the relative early-onset cancer risk is 2.87 times higher among siblings and mothers given a proband with leukemia, and 4.66 times higher given a proband with lymphoma\(^7\). Acute lymphoblastic leukemia (ALL) is the most common leukemia in children and young adults. Recent studies have identified deleterious germline variants in TP53, PAX5, IKZF1, and ETV6 as risk factors for ALL\(^9,10,11\) but have not assessed the spectrum of deleterious germline variants among ancestral groups or in individuals with a family history identified in a population-based registry. Concerning the elevated familial risk of lymphoid malignancies other than leukemia that our group has observed in California\(^7\), previous studies have reported similar results: that first-degree relatives of patients with non-Hodgkin lymphoma (NHL), Hodgkin lymphoma (HL), or chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL) among European origin populations have \(~1.7\) times, 3.1 times, and 8.5 times higher
risk of developing these malignancies, respectively\textsuperscript{12-17}. Rare variants associated with familial-associated lymphoma risk, however, have not been studied extensively\textsuperscript{18}.

The risk of hematologic cancers varies by ancestral/ethnic group. In California, Latino individuals have a higher incidence of ALL compared to those of other ancestral/ethnic groups across all ages\textsuperscript{19}; however, the highest familial risk (identified by families with 2 or more independent sibling cancers) was found among non-Latino Asian/Pacific Islanders (NLAPI)\textsuperscript{7}. For lymphomas diagnosed under 19 years of age, the incidence among non-Latino White (NLW) subjects is higher than subjects of other ancestral/ethnic groups\textsuperscript{7}. However, the variation in germline predisposition that may drive this difference in risk by ancestral/ethnic group has not been formally addressed. Standard US state-based population registries may be useful to study rare genetic predispositions in an agnostic, “real world” population-based manner when biological samples are available. Here, we utilize linked cancer population registries in California along with the California Biobank to identify the nature and ancestral origin of rare P/LP germline variants that may explain this variation in cancer risk.

**Methods**

Our study methods were reviewed and approved by the Committee for Protection of Human Subjects (IRB) of the State of California. Cancer patients were derived from linked population-based registries in California as previously described\textsuperscript{7}. Briefly, the dataset was created by linking information from the California Cancer Registry (CCR) and California Birth Statistical Master File. The linked dataset encompassed all cancer cases comprehensively diagnosed in individuals aged 0-26 years old, since 26 is the oldest age for which it was possible to perform this linkage at the time of this work; their sibling’s; and their mother’s cancers also captured within this registry. To study the rare genetic predispositions in association with hematologic cancers, we first identified all sibling pairs with cancer in the
database. For all individuals with hematologic cancers (leukemia or lymphoma) and a sibling with cancer (of any type, excluding any iatrogenic cancers), we processed the pair for sequencing. Overall, 70 patients from 35 families were eligible for inclusion.

DNA Preparation and sequencing, and mapping and variant identification were performed as explained in online Supplemental Methods. The GATK pipeline for germline short variant calling was performed, with variant filtering as appropriate. Presence of Down syndrome was assessed by chromosome 21 copy number normalized to all other chromosomes. Pathogenicity of variants was assessed using informatic tools (Varsome, PecanPie) initially, followed by strict manual curation of all identified alleles by ACMG/AMP guidelines.

The ancestral identity of each individual was assessed by matching local haplotypes with one of 5 Earth superpopulations (LAT, EUR, EAS, AFR, SAS). The informatic tool RFMix was used for this purpose, and the global ancestry of each individual calculated by summing up the ancestral components of each of the haplotypes for the entire genome.
Results

Demographics of study participants

Among the 70 patients eligible for inclusion, 67 samples had adequate DNA for genotyping. 64 of the 67 samples were sibling pairs from the same family and were included in the genetic analyses. The 64 subjects were from 32 families, among which 12 (38%) were Latino and 10 (31%) were non-Latino White (NLW) families (Table 1). Including parents and all healthy or diseased children, the average size of each Latino family was 4.92 people/family and the average size of NLW families was 6.10 people/family. One out of the 32 families had a mother diagnosed with cancer (desmoplastic small round cell tumor) in her twenties; we note that given the recruitment period of 1989-2015 we have a limited available follow up time to track adult cancer in parents. We do not report age of onset of any cancer case to protect confidentiality of the families evaluated in this study. One subject (#534, family #2) was identified to have DS (# chr21 reads = 917228; population mean and SD of chr21 = 651691 +/- 128812) (Supplementary Tables 3&4).

Pathogenic/Likely Pathogenic and VUS variants

The mean sequencing coverage across subjects was 69.1 (range: 37.2-98.7) (Supplementary Tables 3&4). Our initial software-guided curation yielded 14 unique P/LP variants. Upon further refined manual curation, five P/LP variants were reclassified as VUS, leaving 9 rare P/LP variants shared between siblings among 8 sibling pairs (Supplementary Table 3) and 105 rare VUS that are shared between siblings among the 32 families (Supplementary Table 4). Notably, along with demoting five LP variants to VUS, manual curation rescued on VUS to be LP (HABP2), and reclassified another 23 VUS to be LB (likely benign, Supplementary Table S3). Among the P/LP variants, one (11%) was found in Latino families, 2 (22%) were found in NLW families, 2 (22%) were found in non-Latino Black (NLB) families, 2 (33%) were found in non-Latino Asian/Pacific Islander (NLAPI) families, and 1 (11%)
was found in a non-Latino American Indian/Alaskan Native (NLAIAN) family. Among the VUS, 41 (39%) were found in Latino families, 21 (22%) were found in NLW families, 10 (10%) were found in NLB families, 27 (28%) were found in NLAPI families, and 1 (1%) was found in a NLAIAN family (Table 2). In the Latino families, P/LP variants were detected in *ATM* (NM_000051.3:c.3158dup, one sibling had T-cell leukemia and the other had diffuse large B-cell lymphoma) in family #27; For NLW families, P/LP variants were detected on *TP53* (family #6, NM_000546.5:c.586C>T, one sibling with acute leukemia, not otherwise specified [NOS], and the other with signet ring carcinoma) and *NOD2* (family #25, NM_022162.2:c.1515dup, both siblings diagnosed with Hodgkin lymphoma, NOS). For NLB families, P/LP variants were detected on *GATA2* (family #15, NM_001145661.1:c.1009G>T, one sibling AMML-M4; and second sibling, AMKL-M7) and *NOD2* (family #35, NM_022162.2:c.1515dup, both siblings with Burkitt cell leukemia). In separate NLAPI families, P/LP variants were detected on *FLG* (family #2, NM_002016.1:c.3321del, Hodgkin lymphoma and AMKL-M7), and *CBLB* (family #21, NM_170662.4:c.2307del, large B-cell lymphoma and neuroblastoma). The siblings in this family also shared an additional nonsense variant in *HABP2*. For one NLAIAN family (#29), a P/LP variant was detected on *SBF2* (NM_030962.3:c.4400del, one sibling with acute lymphoblastic leukemia, NOS; and the other sibling with hepatoblastoma). (Supplementary Table 2).

Additionally, we identified 7 rare P/LP variants that are not shared between the two siblings. In 4 Latino subjects without any P/LP variants shared with siblings, these included a nonsense variant in *AMER1* (NM_152424.3:c.28C>T), a frameshift deletion in *HLTF* (NM_003071.3:c.1967del), and missense variants in *GJB2* (NM_004004.5:c.596C>T) and *PINK1* (NM_032409.2:c.1040T>C) (Supplementary Table 5). In one NLW subject diagnosed with a malignant teratoma, we found a non-shared nonsense variant in *SMARCA4* (NM_001128849.1:c.4038G>A). In a NLB subject who developed acute megakaryoblastic leukemia (AMKL), and who also shared a P/LP variant in *GATA2* with their sibling who developed AML (family #15), a non-shared nonsense variant was identified in *FAT1*
(NM_005245.3:c.10957G>T). Finally, in a NLAPI subject who developed AMKL, and shared a P/LP variant in FLG with their sibling who developed Hodgkin lymphoma (family #2), a non-shared rare P/LP variant was identified in exon 2 of GATA1 (NM_002049.3:c.115G>T) (Supplementary Table 5), in which somatic mutations are known drivers of AMKL. Of note, this GATA1 variant had an allele fraction of 71.4% and was detected in the subject identified with trisomy 21 (subject #534, see above in Demographics); thus, it is likely a somatic GATA1 mutation that would be indicative of transient abnormal myelopoiesis (TAM), which occurs frequently in newborns with DS and is a precursor to AMKL in some cases20.

**Ancestry of P/LP variants**

For the one Latino family with a shared sibling variant, the P/LP variant in ATM was shared on a South Asian haplotype. For NLW families, the P/LP variants on TP53 and NOD2 shared EUR ancestry, which match their predicted dominant global ancestry. For one NLB family, the P/LP variant on GATA2 shared AFR ancestry on local haplotypes and this matches their predicted dominant global ancestry. For NLAPI families, two pathogenic variants were classified as AMR and one (the novel HABP2 variant) as EAS. For the one NLAIAN family, the P/LP variant on SBF2 has shared local EUR ancestry, which matched their predicted dominant global ancestry. (Figure 1 & Supplementary Table 3).

**Discussion**

This is the first study to our knowledge using purely population-based, no patient contact resources to examine genetic causes of family-based cancer clustering. We demonstrate that deleterious variants shared between family members contribute to early-onset hematologic cancers and are discoverable using California’s linked birth and cancer registries and neonatal blood spot Biobank. This establishes the utility of these population-based resources to study
familial cancer predisposition successfully as well as to identify germline inherited risk alleles among varied ancestral groups. The profile of germline P/LP variants across ancestral/ethnic groups share similarities as well as distinctive characteristics. Consistent with previous findings, we note that individuals with familial early-onset hematologic cancers harbor deleterious germline variants in TP53, GATA2, and ATM. Pathogenic variants in TP53 have been associated with many hematologic cancers including lymphoblastic, myeloid leukemias, and lymphomas. Deleterious germline variants in GATA2 have been associated with myeloid malignancies, and variants in ATM have been associated with T-cell prolymphocytic leukemia, mantle cell lymphoma, and gliomas. The non-shared SMARCA4 germline P/LP variant that we identified in a subject with teratoma is consistent with previous reports that have described germline mutations in SMARCA4 for teratomas.

Some P/LP variants exhibit novelty. The shared SBF2 in one NLAIAN family is a novel discovery in childhood cancers. SBF2 is associated with cancer by a newly identified long non-coding RNA, SBF2-AS1. SBF2-AS1 was initially characterized in non-small cell lung cancer, and has recently been reported to overexpress in multiple adulthood cancers in East Asian population. P/LP variants in NOD2 is also a novel discovery in childhood cancers, and intriguingly one variant was found in two siblings who both developed Hodgkin lymphoma. NOD2 is a gene that is involved in immune response, and has been associated with colorectal cancer. In addition, we also reported a shared FLG mutation in association with childhood hematologic cancers. Somatic mutations of this gene have been reported to be associated with autoimmune diseases. A HABP2 mutation was shown to be shared between siblings with malignant large-B-cell lymphoma and neuroblastoma. HABP2 encodes a serine protease, and an inactivating mutational variant (G534E) distinct from the one found here was previously linked to extramedullary thyroid neoplasia. The nonsense mutation found here in this gene (C290*) has not been reported previously and likely represents a private familial mutation.
There are several strengths and limitations of this study. A major strength is the linkage of population-based cancer registries to identify subjects for sequencing. Thus, we have established a novel perspective on genetic predisposition that drive the excessive familial risk of early-onset hematologic cancers without selection or recruitment bias that may affect clinic or referral-based studies – and found P/LP alleles in about a quarter (25%) of families as well as many suspect VUS alleles in addition, and an additional 7 families (22%) that have a P/LP mutation in one child only. This compares with the ~5-10% frequency of such alleles in sample series chosen without regards to family history\(^3\)\(^-\)\(^6\). A study that examined predisposition in the context of multiple sibling myeloid cancers in families found candidate mutations in 83 of 86 families, using anectodical clinical referral recruitment of families and liberal mutation identification criteria\(^5\)\(^1\). While we also similarly found VUS and P/LP mutations in nearly every subject we emphasize the reporting of highly curated L/LP alleles – which in our 64 subjects yielded 23 (36%) mutations. Given the large size of our base population and definition of “cancer families” as those with a minimum of only 2 cases, it is likely that some families will be afflicted by multiple cancers by chance without predisposing alleles or be influenced via familial sharing of an environmental exposure. Still, our 36% yield compares with a 4.4% yield in an assessment of leukemia-associated P/LP mutations in a large study of sporadic patients\(^4\); attesting to the power of population-based recruitment in the current investigation.

Benefiting from the highly diverse population in California, we are the first research group that has assessed the variation in genetic predisposition in association with familial hematologic cancer risk in multiple ancestral/ethnic groups over a defined population, appraising cancer predisposition in a selection-bias free, consent-bias free population. However, limited by the length of time in data collection and the rarity of childhood cancers, we have only identified a selection of mutations that are unique to an ancestral/ethnic group. In addition, we have no way to evaluate or confirm the impact of nonshared P/LP alleles inherited by chance by those siblings that did not share a P/LP allele. Also, limited to analysis of SNVs and indels, we
did not assess variation in copy number or epigenetic alterations that may be potentially pathogenic.

Cancer risk is increased by both genetic and environmental factors. Family members are generally exposed to the same environments and therefore those in-common environments may contribute to the observed familial cancer clustering. To understand such clustering better, concomitant assessments of environmental factors in coordination with germline genetics should be considered for future studies. Such analyses will help to establish the role of strong cancer predisposition variants more definitively among different ancestral/ethnic groups, and whether such mutations harbor varied penetrance in the context of different ancestral backgrounds and environments. This type of evaluation is increasingly critical in the intermixed population of California and other world regions and with profound implications for family genetic counseling in the context of improved cancer patient survival.
References
Table 1. Demographics of the 32 proband-sibling pairs, California Cancer Registry, 1989-2015.

<table>
<thead>
<tr>
<th></th>
<th>Overall (N=64)</th>
<th>Proband (N=32)</th>
<th>Affected sibling (N=32)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-reported race/ethnicity, N (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latino all races</td>
<td>24 (37.5%)</td>
<td>12 (37.5%)</td>
<td>12 (37.5%)</td>
</tr>
<tr>
<td>NLW</td>
<td>20 (31.3%)</td>
<td>10 (31.3%)</td>
<td>10 (31.3%)</td>
</tr>
<tr>
<td>NLB</td>
<td>6 (9.4 %)</td>
<td>3 (9.4%)</td>
<td>3 (9.4%)</td>
</tr>
<tr>
<td>NLAPI</td>
<td>12 (18.8%)</td>
<td>6 (18.8%)</td>
<td>6 (18.8%)</td>
</tr>
<tr>
<td>NLAIAN</td>
<td>2 (3.1%)</td>
<td>1 (3.1%)</td>
<td>1 (3.1%)</td>
</tr>
<tr>
<td><strong>Cancer site, N (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leukemias</td>
<td>27 (42.2%)</td>
<td>12 (37.5%)</td>
<td>15 (46.9%)</td>
</tr>
<tr>
<td>Lymphomas</td>
<td>17 (26.6%)</td>
<td>9 (28.1%)</td>
<td>8 (25.0%)</td>
</tr>
<tr>
<td>Brain tumors</td>
<td>5 (7.8%)</td>
<td>3 (9.4%)</td>
<td>2 (6.2%)</td>
</tr>
<tr>
<td>Neuroblastoma</td>
<td>1 (1.6%)</td>
<td>1 (3.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Retinoblastoma</td>
<td>1 (1.6%)</td>
<td>1 (3.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Renal tumors</td>
<td>1 (1.6%)</td>
<td>1 (3.1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Hepatic tumors</td>
<td>1 (1.6%)</td>
<td>0 (0%)</td>
<td>1 (3.1%)</td>
</tr>
<tr>
<td>Bone tumors</td>
<td>1 (1.6%)</td>
<td>0 (0%)</td>
<td>1 (3.1%)</td>
</tr>
<tr>
<td>Sarcomas</td>
<td>5 (7.8%)</td>
<td>3 (9.4%)</td>
<td>2 (6.2%)</td>
</tr>
<tr>
<td>Germ cell tumors</td>
<td>3 (4.7%)</td>
<td>2 (6.2%)</td>
<td>1 (3.1%)</td>
</tr>
<tr>
<td>Neoplasms and melanomas</td>
<td>2 (3.1%)</td>
<td>0 (0%)</td>
<td>2 (6.2%)</td>
</tr>
<tr>
<td><strong>Sex of child, N (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>38 (59.4%)</td>
<td>16 (50.0%)</td>
<td>22 (68.8%)</td>
</tr>
<tr>
<td>Female</td>
<td>26 (40.6%)</td>
<td>16 (50.0%)</td>
<td>10 (31.2%)</td>
</tr>
<tr>
<td><strong>Age, mean (SD)</strong></td>
<td>9.75 (7.10)</td>
<td>6.22 (5.45)</td>
<td>13.3 (6.86)</td>
</tr>
</tbody>
</table>

Abbreviations: NLW, non-Latino White; NLB, non-Latino Black; NLAPI, non-Latino Asian/Pacific Islander; NLAIAN, non-Latino American Indian/Alaskan Native.
Table 2. Distribution of rare variants shared by two siblings in a family, California Cancer Registry, 1989-2015.

<table>
<thead>
<tr>
<th>Pathogenic variants (N=9)</th>
<th>VUS (N=105)</th>
<th>Overall mutations (N=113)</th>
<th>No. of families (N=32)</th>
<th>VUS per family</th>
<th>Overall mutation per family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reported race/ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latino all races</td>
<td>1 (11)</td>
<td>41 (39)</td>
<td>44 (39)</td>
<td>12 (38)</td>
<td>3.2</td>
</tr>
<tr>
<td>NLW</td>
<td>2 (22)</td>
<td>21 (22)</td>
<td>24 (21)</td>
<td>10 (31)</td>
<td>2.2</td>
</tr>
<tr>
<td>NLB</td>
<td>2 (22)</td>
<td>10 (10)</td>
<td>12 (11)</td>
<td>3 (9)</td>
<td>3.3</td>
</tr>
<tr>
<td>NLAPI</td>
<td>2 (22)</td>
<td>27 (28)</td>
<td>31 (27)</td>
<td>6 (19)</td>
<td>4.7</td>
</tr>
<tr>
<td>NLAIAN</td>
<td>1 (11)</td>
<td>1 (1)</td>
<td>2 (2)</td>
<td>1 (3)</td>
<td>1.0</td>
</tr>
<tr>
<td>Category</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>6 (73)</td>
<td>79 (68)</td>
<td>78 (69.0)</td>
<td>28 (58)</td>
<td>2.4</td>
</tr>
<tr>
<td>Hematological</td>
<td>1 (7)</td>
<td>10 (11)</td>
<td>12 (10.6)</td>
<td>11 (23)</td>
<td>1.0</td>
</tr>
<tr>
<td>Immunological</td>
<td>1 (13)</td>
<td>9 (9)</td>
<td>11 (9.7)</td>
<td>8 (17)</td>
<td>1.1</td>
</tr>
<tr>
<td>Other</td>
<td>0 (0)</td>
<td>2 (2)</td>
<td>1 (0.9)</td>
<td>1 (2)</td>
<td>1.0</td>
</tr>
<tr>
<td>Mutation class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frameshift</td>
<td>6 (67)</td>
<td>6 (5)</td>
<td>6 (14)</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Missense</td>
<td></td>
<td>105 (97)</td>
<td>101 (89)</td>
<td>30 (71)</td>
<td>3.2</td>
</tr>
<tr>
<td>Nonsense</td>
<td>3 (33)</td>
<td>0 (0)</td>
<td>3 (3)</td>
<td>3 (7)</td>
<td>0.3</td>
</tr>
<tr>
<td>Protein deletion</td>
<td>2 (2)</td>
<td>2 (2)</td>
<td>2 (5)</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Splice</td>
<td>1 (1)</td>
<td>1 (2)</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: VUS, variants of uncertain significance; NLW, non-Latino White; NLB, non-Latino Black; NLAPI, non-Latino Asian/Pacific Islander; NLAIAN, non-Latino American Indian/Alaskan Native; PeCanPIE, Pediatric Cancer Variant Pathogenicity Information Exchange.

1 p-value calculated with a Chi-squared test to compare the difference in VUS per family among subjects of each race/ethnic group to the VUS per family among Latino subjects.

2 p-value calculated with a Chi-squared test to compare the difference in overall mutations per family among subjects of each race/ethnic group to the overall mutations per family among Latino subjects.
Figure legend

Figure 1. Predicted global ancestries of study subjects and local ancestries of putatively pathogenic variants, California Cancer Registry, 1989-2015. *Predicted local ancestry matches the predicted dominant global ancestry. Abbreviations: LW, Latino White; NLW, non-Latino White; NLB, non-Latino Black; NLAPI, non-Latino Asian/Pacific Islander; NLAIAN, non-Latino American Indian/Alaskan Native; EUR, European; AFR, African; AMR, Admixed American; EAS, East Asian; SAS, South Asian.
<table>
<thead>
<tr>
<th>Family ID</th>
<th>Subject ID</th>
<th>Global Ancestry (percentage)</th>
<th>Race/Ethnicity (self report)</th>
<th>Gene</th>
<th>Local_hap1</th>
<th>Local_hap2</th>
<th>Mutation shared on</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>369-435</td>
<td></td>
<td>LW</td>
<td>ATM</td>
<td>SAS</td>
<td>EUR</td>
<td>SAS</td>
</tr>
<tr>
<td>6</td>
<td>277-396</td>
<td></td>
<td>NLW</td>
<td>TP53</td>
<td>AFR</td>
<td>EUR</td>
<td>EUR</td>
</tr>
<tr>
<td>25</td>
<td>203-287</td>
<td></td>
<td>NLW</td>
<td>NOD2</td>
<td>EUR</td>
<td>EUR</td>
<td>EUR</td>
</tr>
<tr>
<td>15</td>
<td>33-269</td>
<td></td>
<td>NLB</td>
<td>GATA2</td>
<td>EAS</td>
<td>AFR</td>
<td>AFR</td>
</tr>
<tr>
<td>35</td>
<td>28-610</td>
<td></td>
<td>NLB</td>
<td>NOD2</td>
<td>EUR</td>
<td>EUR</td>
<td>Unknown</td>
</tr>
<tr>
<td>2</td>
<td>412-534</td>
<td></td>
<td>NLAPI</td>
<td>FLG</td>
<td>AMR</td>
<td>EUR</td>
<td>AMR</td>
</tr>
<tr>
<td>21</td>
<td>297-302</td>
<td></td>
<td>NLAPI</td>
<td>CBLB</td>
<td>SAS</td>
<td>AMR</td>
<td>AMR</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td>HABP2</td>
<td>EAS</td>
<td>EAS</td>
<td>EAS</td>
</tr>
<tr>
<td>29</td>
<td>258-364</td>
<td></td>
<td>NLAIAN</td>
<td>SBF2</td>
<td>EUR</td>
<td>EUR</td>
<td>EUR</td>
</tr>
</tbody>
</table>
Supplemental 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. (see excel files)

Supplementary Table 1. List of genes evaluated for pathogenicity.

Supplementary Table 2. 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.

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

Supplementary Table 4. Variants of uncertain significance (VUS) shared between family members.

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

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\textsuperscript{3-5}, based on the GRCh37 assembly. Resulting gene sequence variations stored in variant call format (VCF) files were annotated with ANNOVAR\textsuperscript{6}. Variants with alternative allele reading depth $\leq 5$, or variant allele fraction $\leq 0.2$ 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)\textsuperscript{7,8} and the 1000 Genomes Project (1KG)\textsuperscript{9}. Rare variants with both global and population-specific allele frequency $\leq 0.001$ 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)\textsuperscript{10}/Association for Molecular Pathology (AMP) guidelines\textsuperscript{11}. The potential pathogenicity of the variants is classified into three tiers (gold, silver, and bronze)\textsuperscript{12}. We examined variants on a list of 1173 genes that are reported to be cancer/immunodeficiency/nonmalignant hematological-related genes (n=986)\textsuperscript{12,13}, pediatric...
cancer predisposition genes (n=162)\textsuperscript{14, 15}, tumor suppressor genes, tyrosine kinase genes, or cancer genes classified based on their recurrent somatic mutation in cancer (n=565)\textsuperscript{16}, and/or Hodgkin lymphoma-related genes identified from genome-wide association or sequencing studies (n=327)\textsuperscript{17-25} (Supplementary Table 1).

Then, the pathogenicity of ‘gold’ and ‘silver’ medal variants identified by PeCanPIE was cross-checked with VarSome\textsuperscript{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 \textit{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 Supplemental Table 2 only; all manually curated alleles are displayed in other key tables (Table 2, Figure 1, Supplemental Table 3, 4, 5).

All putative P/LP variants and VUS that were shared by both siblings were inspected visually with the Integrative Genomics Viewer\textsuperscript{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.

\textit{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\textsuperscript{28} was used to determine the global and local ancestries.

We used gene sequence variations from the 1KG\textsuperscript{9} and Human Genome Diversity Project (HGDP)\textsuperscript{29} to construct a reference panel for the ancestry analysis with RFMix. First, ADMIXTURE\textsuperscript{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.


