Large-Scale Identification of Clonal Hematopoiesis and Mutations Recurrent in Blood Cancers

Julie E. Feusier1,2,*, Sasi Arunachalam1,3,4,*, Tsewang Tashi3,5,6, Monika J. Baker1, Chad VanSant-Webb1, Amber Ferdig1, Bryan E. Welm3,7, Juan L. Rodriguez-Flores8, Christopher Ours1, Lynn B. Jorde2, Josef T. Prchal3,5,6, Clinton C. Mason1,ξ

1Department of Pediatrics, Division of Pediatric Hematology and Oncology, University of Utah, Salt Lake City, UT, USA
2Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
3Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA
4Department of Oncological Sciences, University of Utah, Salt Lake City, UT, USA
5Division of Hematology and Hematologic Malignancies, University of Utah, Salt Lake City, UT, USA
6VA Medical Center, Salt Lake City, UT, USA
7Department of Surgery, University of Utah, Salt Lake City, UT, USA
8Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, USA
*These authors contributed equally to this work
ξCorresponding Author: Clinton C. Mason, PhD, 417 Wakara Way, Room 3117, Salt Lake City, Utah 84108; Ph: 801-587-3633; E-Mail: clint.mason@hsc.utah.edu

Word Count: 3,498
Figure/Table Count: 3/1
The authors declare no conflict of interests.

Running title: Clonal hematopoiesis and recurrent blood cancer mutations

Keywords – clonal hematopoiesis of indeterminate potential, recurrent mutation, mutational hotspot, oncogenesis, pre-cancer surveillance
Abstract

Clonal hematopoiesis of indeterminate potential (CHIP) is characterized by detectable hematopoietic-associated gene mutations in a person without evidence of hematologic malignancy. We sought to identify additional cancer-presenting mutations useable for CHIP detection by performing a data mining analysis of 48 somatic mutation studies reporting mutations at diagnoses of 7,430 adult and pediatric patients with hematologic malignancies. Following extraction of 20,141 protein-altering mutations, we identified 434 significantly recurrent mutation hotspots, 364 of which occurred at loci confidently assessable for CHIP. We then performed an additional large-scale analysis of whole exome sequencing data from 4,538 persons belonging to three non-cancer cohorts for clonal mutations. We found the combined cohort prevalence of CHIP with mutations identical to those reported at blood cancer mutation hotspots to be 1.8%, and that some of these CHIP mutations occurred in children. Our findings may help to improve CHIP detection and pre-cancer surveillance for both children and adults.
Significance

This study identifies frequently occurring mutations across several blood cancers, which may drive hematologic malignancies and signal increased risk for cancer when detected in healthy persons. We find clonal mutations at these hotspots in a substantial number of individuals from non-cancer cohorts, including children, showcasing potential for improved pre-cancer surveillance.
Introduction

Somatic mutations at hotspots (genetic loci observed to be frequently mutated across cancer patients) often drive or contribute to cancer pathogenesis (1,2). Many somatic landscape studies have now been performed with next generation sequencing (NGS), typically for a single cancer type in small- to medium-sized patient cohorts, identifying mutations that recur frequently. Large single study and pan-cancer analyses of both pediatric and adult cohorts have identified cancer genes and mutation hotspots having even lower but significant recurrence rates. For example, over 140 driver genes were identified in an analysis of children with six cancer types (3) and more than 450 hotspots were found in an analysis of 41 predominantly nonhematologic cancers (2). However, the number of included leukemic or other hematological cancer samples in both pan-cancer and single study analyses reporting hotspots at a codon level has generally been insufficient to identify blood cancer-specific hotspots of low recurrence. Existing databases that accumulate mutations across many individual studies (most notable in size, the Catalogue of Somatic Mutations in Cancer (COSMIC, cancer.sanger.ac.uk) (4)) are extremely valuable for identifying reported mutations within their included studies; yet current versions of these databases lack many exome/genome-wide blood cancer studies that could improve detection of mutations recurrent at low frequencies, and some are difficult to filter for specific cancers. Hence, many hematologic malignancies’ mutation hotspots of lesser recurrence and their constituent mutations that may drive blood cancers are likely still unidentified or unappreciated.

Clonal hematopoiesis of indeterminate potential (CHIP), which is detected as an expansive clonal somatic mutation in a person currently free from hematologic malignancy, carries a greatly increased risk (hazard ratio >10) for future blood cancer development (5,6). CHIP increases sharply in prevalence with advanced age (5–9), and backwards extrapolation of trends...
in adults could suggest that CHIP in childhood is extremely rare. However, no exome-wide CHIP study has included large numbers of healthy children to directly address this issue. The mutations in hematopoiesis-regulating genes used to identify CHIP include hotspot mutations of high recurrence, but also include mutations lacking driver or recurrent evidence at specific amino acid positions. Regardless of neoplastic evidence, CHIP mutations are frequently identified in persons singly (5–10); therefore, most are likely responsible for the clonal expansion that makes their detection possible. Yet CHIP mutations identical to those reported at blood cancer mutation hotspots may be found to carry increased risk for future neoplastic transformation.

With the dual aims of identifying mutational hotspots of hematologic malignancies and using these hotspots to detect CHIP having potentially greater risk for affecting cancer transformation, we performed a recurrent mutation analysis independently from existing databases. We systematically extracted and analyzed reported data from 48 studies meeting inclusion criteria, which detected mutations at diagnoses of patients with acute leukemias, myeloproliferative neoplasms (MPNs), and myelodysplastic syndrome (MDS). As the majority of these studies interrogated the entire exome and many are not currently included in other mutation databases, our analysis provides a valuable compiled list of recurrent hematologic cancer mutations at a protein-coding level with smaller genome-wide bias and larger sample size than previously reported. We then performed a large analysis of CHIP, focused on finding clonal mutations identical to those reported at these hotspots in the whole exome sequencing (WES) data of 4,538 persons from three non-cancer cohorts, which included the widely used 1000 Genomes Project and more than 400 children. Our findings may lead to the improved prognostic and clinical utility of pre-cancer surveillance in children and adults.
Results

Somatic mutation studies.

By systematic literature search, we identified 48 somatic mutation landscape studies of blood cancers (3,11–57) that could be used to determine recurrence of mutations reported at amino acid positions in protein-coding genes (Table 1). Focusing on diseases which may have been preceded by CHIP (58), we limited our investigation to studies which assessed patients by NGS at diagnosis of one of seven hematologic malignancies: acute lymphoid leukemia (ALL), acute myeloid leukemia (AML), chronic myeloid leukemia (CML), chronic myelomonocytic leukemia (CMML), juvenile myelomonocytic leukemia (JMML), MDS, and MPN (Fig. 1). After evaluation, filtering, and harmonization of 58,177 reported mutations assessed in 7,430 diagnostic patients, we determined a total of 20,141 mutations that altered an amino acid or splice site.

Mutational hotspots in hematologic cancers.

We found 434 amino acid or splice-site positions occurring across 85 genes that met our criteria for hotspots (Supplementary Table S1). The most common hotspots were well-known and observed frequently even within individual studies, (e.g., mutations at NRAS p.G12 were reported for 345 persons within 35 studies, Fig. 2A). Yet, 79 hotspots observed in three to eight persons had only a single mutation within any given study, highlighting the benefit of multi-study evaluations. Nearly all of the hotspots were recurrently listed across combined cancers in the most recent COSMIC database which includes many blood cancer studies, the majority having been published between 1992 and 2010. Yet this version was lacking 32 of the 48 NGS
studies (comprising 51% of total patients) used in our assessment. When restricting COSMIC mutation data to primary samples in the seven hematologic malignancies of our focus, the majority (n=222) of our identified hotspots were observed in less than three individuals of this COSMIC subset (Supplementary Table S2).

Some of our identified mutation hotspots were unique to a specific hematologic malignancy (Fig. 2B; Supplementary Table S1). With the proportion of assessed patients being 49.4% (AML), 31.0% (ALL), and 19.6% (other malignancies), we found that mutations at IDH2 p.R172 (n=59), KIT p.N822 (n=31), and KIT p.Y418 (n=18) were reported only in AML patients, whereas RPL10 p.R98 (n=25), NOTCH1 p.L1678 (n=24), FBXW7 p.R479 (n=23), NOTCH1 p.L1600 (n=22), and NOTCH1 p.L1585 (n=19) mutations were only observed in patients with ALL. We also observed that some genes contain mostly point or non-frameshift hotspots (e.g., KRAS, NRAS, PTPN11, SF3B1), nonsense or frameshift hotspots (e.g., ASXL1, NF1, STAG2), or had distinct groupings of mutation types by location within the coding region (e.g., CEBPA, NOTCH1).

**Hotspot mutations evaluable for CHIP.**

We identified 70 of the 434 hotspots to have heightened potential for false positives if used to infer CHIP from mutations in blood samples alone (Supplementary Table S3). This resulted in 364 hematologic cancer mutation hotspots and 755 constituent mutations that could be confidently used in screening for CHIP with prevalent NGS methods (Supplementary Table S4). Only 45 (12.4%) of these 364 loci were identified as hotspots in a previous pan-cancer analysis (2) and only 35 (9.6%) were observed in three or more patients of a sizeable MDS investigation.
not included in our assessment (59). Likely because many of the exome-wide studies that identified mutations at these hotspots were published around or after a previous landmark CHIP study (10), 350 of the 755 specific mutations identified in our hotspot mutation assessment were not present in the list of queried CHIP variants of that study. Of these, 134 were reported in ≥3 patients at diagnosis of a hematologic malignancy (Supplementary Table S5). Many of these unique mutations occur within \textit{NOTCH1}, \textit{FLT3}, \textit{CEBPA}, \textit{KIT}, and \textit{RUNX1}.

**CHIP in non-cancer cohorts identical to hotspot mutations.**

We next assessed three large non-cancer cohorts (60–62) for CHIP at our identified hotspots within their combined 4,538 individuals. The 1000 Genomes Project (1KG) cohort included 2,503 adults from 26 different populations with sample sizes ranging from N=61 to N=113 (median N=99). The Qatari Genome (QTRG) and Simons Simplex Collection (SSC) cohorts included children and adults. We found the ages, sequencing depths, and methods utilized in each study to vary considerably (Supplementary Fig. S1A), with potential to influence the rate of clonal mutation identification. The means of the average depths of coverage across the 364 confident hotspots were 78.5x, 93.3x, and 53.6x with averages of 27.5%, 32.7%, and 10.3% of hotspots covered at a depth ≥100x in the 1KG, SSC, and QTRG cohorts respectively (Supplementary Fig. S1B).

All detected variants were evaluated for reliability (which led to the exclusion of 8 outlier samples) and for meeting the criteria for CHIP at hotspots, resulting in the identification of 83 individuals (1.83% of the 4,530 included) each having a single CHIP mutation at one of 62 hotspots across 23 genes (Supplementary Fig. S2; Supplementary Table S6). The prevalence rate
of CHIP at hotspots for each cohort was: QTRG: 0.73% (9/1,231), 1KG: 2.48% (62/2,503), and SSC 1.51% (12/796) – with subgroup rates of 0.77% (3/388) for SSC children and 2.21% (9/408) for SSC adults (Fig. 3A). CHIP mutations at hotspots were most common in DNMT3A, TET2, and TP53 (Fig. 3B). Clonal size was not significantly associated with reported frequency in our recurrent mutation analysis (p=0.56, Supplementary Fig. S3). However, CHIP detection was associated with frequency of hotspot recurrence (odds ratio, 1.75 per log of reported mutations; 95% confidence interval [CI], 1.33 to 2.29; p<0.0001). Still, while half of the 12 hotspots (including those of lesser confidence for CHIP evaluation) reported in more than 100 patients of the recurrence analysis were observed with identical clonal mutations in the non-cancer cohorts, six were not. These included NPM1 p.W288 and FLT3 p.D835, both of which may rapidly accelerate neoplastic onset, making them more difficult to observe in healthy persons (9). Together, these observations suggest that stratifying CHIP mutations by significant recurrence in hematologic malignancies may prove prognostic in studies of blood cancer risk.

**CHIP at mutation hotspots by age-groups.**

We detected CHIP mutations at the leukemic hotspots in 3/388 (0.77%) children of the SSC cohort (two autism spectrum disorder probands and one unaffected sibling). These mutations occurred at hematologic hotspots in DNMT3A (p.R882S, variant allele frequency (VAF) = 2.5%) and RUNX1 (p.G170 splice site (c.509-1G>T), VAF = 1.3%; p.S322X, VAF = 1.1%), with all variant reads passing manual review and being detected on both strands (Supplementary Table S6, Supplementary Fig. S4). We assessed the parent-child trios for each of these children for expected SNP inheritance (to preclude the possibility of sample mix-up of children and adult samples or data) and confirmed that all three samples were from the child. The child with a
RUNXI p.S322X clonal mutation also had an apparent germline SNP at RUNXI p.L56S (variant present in 5/10 reads) that was inherited from the child’s father. This SNP is present in the gnomAD version 2.1.1 database (63) at a frequency of 1.2% yet was also reported in 4/56 (7.1%) patients diagnosed with CMML (46), suggesting a possibility that this variant helped accelerate clonal growth under the two-hit hypothesis (64). Still, the clonal size and number of variant reads at each hotspot in these children were insufficient to completely rule out the possibility of artifacts, and hence these findings are preliminary.

The QTRG cohort having participants with age ranging from 0-85 years had a much lower sequencing depth, so we explored the effect of relaxing the requirement for detecting CHIP at hotspots to include mutations with only two variant reads. While such a threshold would result in a high false positive rate in deeply sequenced data, a higher rate of true positives is likely to be present in lower depth data. This assessment found an additional 41 clonal mutations for this cohort (Supplementary Table S7) and found age to be associated with relaxed CHIP prevalence in this call-set (odds ratio, 1.025 per year of age; 95% CI, 1.004 to 1.048; p=0.022; Fig. 3C).

**CHIP not exclusive to hotspots.**

We also investigated the non-cancer cohorts for the prevalence of CHIP that included seldom or never-before reported mutations at diagnosis of blood cancers, but which may still be involved with clonal expansion due to their occurrence in genes with known hematopoietic function. Across the specified mutations and domains of the 74 allowed hematologic genes of Jaiswal et al. (10), and using their criteria, we observed 189 mutations not exclusive to hotspots within the 4,530 persons of our analyzed cohorts (Supplementary Table S8), 40 of which (21.2%) were
identical to mutations reported at confident hotspots in 3 or more diagnostic patients in the 48 somatic landscape publications we assessed (i.e., those listed in Supplementary Table S4). For comparison, we also determined that 76 of the 224 (33.9%) and 298 of the 805 (37.1%) reported CHIP mutations in Jaiswal et al. (5) and Jaiswal et al. (10) respectively were identical to mutations in our hotspot list.

Our general CHIP analysis identified 6 persons having two mutations (5 of whom had a mutation of \textit{DNMT3A}), none with more than two, and the remaining 177 persons having only a single CHIP mutation – yielding a general CHIP prevalence rate of 4.04% (183/4,530) across the cohorts. Including identified clonal mutations having a VAF >4% at the additional hotspots not previously utilized by Jaiswal et al. (10), 193/4,530 (4.26%) persons had qualifying CHIP. As in previous studies, CHIP not exclusive to hotspots was most frequent for \textit{TET2} and \textit{DNMT3A} (Supplementary Table S8). However, we found mutations in \textit{SETD2}, \textit{EP300}, and \textit{KMT2A/D} to be more common in our cohort while \textit{ASXL1} and \textit{JAK2} mutations were noticeably less frequent. Variability of gene prevalence for CHIP mutations in \textit{DNMT3A}, \textit{TET2}, \textit{ASXL1}, and most strikingly \textit{JAK2} between control and cardiovascular disease cohorts was previously observed (10). As the cohorts we assessed likely had fewer underlying cases of cardiovascular disease and younger average age than other CHIP studies using WES data, some of our gene prevalence rates may reflect these differences. Further exploration identified seven \textit{ASXL1} variants having a VAF >2%, yet only one reached the threshold of 4% required for this general CHIP analysis. No additional \textit{JAK2} variants with three or more supporting reads were observed at any frequency.

Potential CHIP at non-hotspots with a VAF >4% was observed in only four children having variant reads detected on both strands (Supplementary Table S8). In one child, the variant (\textit{DNMT3A} p.V665L) had been previously identified and reported as a de novo mutation by Lim.
et al. (65), though it could possibly have been a CHIP mutation that had reached near-complete saturation (the VAF was 47.3%, see Materials and Methods). This particular variant was not observed in 55 individuals having Tatton-Brown-Rahman syndrome, a congenital condition due to germline $DNMT3A$ mutations (66), one of whom developed AML in childhood. Yet the clinical association of autistic spectrum disorder in 20 (36%) of their study participants could be related to this detection of a $DNMT3A$ mutation in an SSC proband. Also intriguing, another of the four children with CHIP at non-hotspots was one of the three children we initially identified to have a hotspot CHIP mutation, making the child’s non-hotspot $TET2$ p.A1863S variant (VAF = 6.5%) a potentially synergistic CHIP mutation with their previously mentioned $RUNX1$ p.G170 splice site hotspot variant (Supplementary Fig. S4). This observation may once again reflect an increased likelihood of observing CHIP in children when two genetic insults are present, as without an additional factor, single clonal mutations in younger persons may have lacked sufficient time to expand to detectable levels.

**DISCUSSION**

Owing to the work of many researchers who published the findings of their somatic mutation studies (3,11–57), we were able to compile a novel list of recurrent hematologic cancer mutations. The majority of these NGS studies assessed the entire exome/genome, allowing for a less biased and larger-scale assessment of diagnostic mutation recurrence in blood cancers than has previously been available. While significant recurrence alone does not imply a driver or even contributory role in cancer development, overall, the hotspots we identified will likely be enriched for pathogenic blood cancer mutations. Importantly, these identified hotspots also
increase the number of mutations having documented recurrence in primary hematological malignancies that can now be used to identify CHIP.

Using these recurrent mutation loci, we analyzed three large non-cancer cohorts, finding clones with mutations identical to those observed at blood cancer mutation hotspots in 1.83% of persons across the combined cohorts. Without restricting to hotspots, our estimate of CHIP prevalence coincides with that of previous WES studies of CHIP, 4-5%, with the proportion of CHIP identical to diagnostic mutations at hotspots ranging from 21% to 37% within these studies. As the vast majority of all persons with identified CHIP are found to have only a single clonal mutation in hematopoiesis-regulating genes, most detected hotspot or non-hotspot mutations have likely driven the clonal expansion. Future studies of large size and long follow-up will be required to determine the degree to which CHIP at hotspots may carry an increased risk for aiding neoplastic transformation, as well as for apparently unrelated nonhematological cardiovascular complications of CHIP (10).

Our analysis of CHIP included for the first time a large assessment of children free from blood disorders (please note that CHIP has been observed in children having aplastic anemia (67) and that postzygotic mutations of generally higher mutation frequencies in children have also been previously assessed (65)). Of great interest, we detected CHIP at mutation hotspots in three of the children in our cohorts. This novel, preliminary detection of CHIP in children unselected for blood disorders is important, as while only a small number of children were observed to have these clonal mutations, this rate may be higher than anticipated from past analyses of adult cohorts (e.g., only 1/1,039 adults aged 20-39 years was detected with CHIP in a previous analysis of WES data) (5). The well-designed Simons Simplex Collection study sequenced children and parental samples simultaneously at nearly identical depths (60). CHIP at hotspots in
these children was detected at 35% of the overall parental frequency (2.2%), encouraging future studies to investigate this phenomenon more thoroughly in children.

Abelson et al. (7) found that AML mutations reported with higher frequency in the COSMIC database (4) were also observed as CHIP more frequently. We similarly found an association between frequency of reported diagnostic mutations across the seven hematologic malignancies and their detection as clonal mutations in the non-cancer cohorts, strengthening the conjecture that mutations at hotspot loci may increase risk for affecting transition from CHIP to hematologic cancer (6). Some of the hotspot mutations we list may have been overlooked in previous targeted designs for CHIP assessment. These mutations may now receive heightened priority when clinically observed and receive further investigation for functional significance and therapeutic potential.

Much remains to be discovered regarding most individual CHIP mutations, including their prevalence by ethnicity, risk for onset of nonhematologic diseases, and concomitant factors associated with their size and clonal dynamics. Using high-depth WGS and RNAseq, future work may expand to include the identification of clonal fusions in CHIP prevalence and risk assessments. While previously thought to be pertinent to adults only, our analysis of non-cancer cohorts identified CHIP at both hotspots and non-hotspots in children, indicating a need to further explore CHIP in younger-age cohorts. Our hematologic cancer-focused hotspot list will allow for subgroup analyses of CHIP in future studies, with promise to improve its prognostic performance and future blood cancer prevention research efforts.

Materials and Methods
**Somatic mutation studies**

An initial feasibility exercise was performed on 14 somatic mutation landscape studies to discover possible extraction-related problems and necessary steps to confidently identify and harmonize reported mutations useable for hotspot determination. This preliminary work was supplemented with a formal PubMed search on studies published before July 2018 with the search criteria: (“genomic landscape”, “somatic landscape”, “genomic profile”, “mutational landscape”, OR “mutation landscape”) AND (“leukemia” OR “leukaemia”). This search returned a total of 172 papers, of which 61 were found to be review articles reporting on no new patients. We assessed studies focused on any of seven disorders: ALL, AML, CML, CMML, JMML, MPN, and MDS.

**Mutation evaluation**

Each study was assessed for whether hematologic neoplasms had been investigated with NGS and for providing information that could accurately identify the genetic position of each alteration and the effect on protein coding. Studies that did not utilize NGS or reported mutations that could not be determined without ambiguity (owing to multiple transcripts for many genes, mutations from studies providing only an amino acid substitution without either an accompanying transcript identifier or genomic positions were generally deemed ambiguous) were not included. For remaining studies, all substitutions and all deletions were assessed separately for provided reference alleles to determine the combination of factors that were used in reporting mutations, specifically, cDNA or gDNA, 0-bp or 1-bp coordinate system, and human genome reference version. If only a trivial number of reference allele discrepancies were observed, these were discarded and the remaining variants were incorporated – otherwise the entire group of substitutions, indels, or entire study was discarded. While translocations and large
structural rearrangements are common for many cancers (68), these were not included in this investigation as our primary purpose was the identification of recurrent events occurring entirely within exonic regions, so that these could be queried for CHIP in WES data (in addition, less than one third of the studies provided breakpoints of large structural rearrangements at a base pair level). Many somatic landscape studies reported having performed validation of their listed mutations; uniformly, we relied on the mutations as reported by the original authors rather than requesting the original NGS data for re-analysis. Overall, our approach was designed to enrich for reported mutations having high confidence for accurate extraction rather than to collect increased numbers of mutations with less overall confidence in their authenticity.

**Mutation filtering and harmonization**

In studies that included mutations from relapse or secondary leukemia samples in addition to primary samples, mutations in the primary diagnostic samples were extracted only if they could be clearly distinguished from those in relapse or secondary leukemia samples. If there was no distinction, the entire study was excluded. For a few studies assessing samples that had been investigated in previous studies, mutations from patients reported in the older analysis were not included. Mutations in genes known for having high false positive rates were frequently filtered by authors, and we likewise removed mutations reported in such genes from studies that had not performed such filtering. Similarly, a few studies reported all observed variants with little or no read/frequency filtering for mutation calling, and variants having minor variant allele frequencies were removed. Finally, a single RefSeq transcript was determined for each gene based on its study-utilized frequency and length, – and each mutation was assessed for its effect on this transcript (listed in Supplementary Table S1). Only splice-site or protein-altering mutations were retained. A total of 48 studies (3,11–57) satisfying all of these stringent criteria were
included in the hotspot mutation analysis. Comparison with COSMIC data was performed with the most recent COSMIC database (version 92, August 2020).

**Hotspot determination**

Similar to Chang* et al.* (2), we assessed mutational recurrence at amino acid and splice site positions in protein-coding genes, yet also including indels in addition to substitutions. The number of reported protein-altering and splice site substitutions and indels across the final set of 48 somatic mutation landscape studies was tallied at each amino acid locus, with indels tallied at the locus at which they began and splice-sites tallied separately. As most studies did not provide silent or non-exonic mutations in their lists, a formal driver analysis was not possible, which may be considered a weakness of this work. Hence, we sought to determine minimalist criteria for our designation of hotspot based on deviance from expected recurrence in binomial distribution modeling that utilized the set of reported protein-altering mutations (Supplementary Fig. S5). Various models were assessed across which a threshold of at least three times recurrence was markedly pronounced for deviation from expectation. Thus, our minimalist criteria for hotspot designation was defined as three or more patients having such mutations at the same amino acid or splice-site position, with 434 loci meeting that criteria. To determine hotspots that could be confidently used in identifying CHIP, we additionally determined the mappability of the genetic sequence surrounding each locus with multiple approaches including an extensively repeated sequence search. This allowed us to exclude hotspots that occurred at difficult to map regions in CHIP analyses. Reported mutations at regions of homopolymer repeats or identical to non-rare germline SNPs were also classified as less confident loci for CHIP calling and excluded. These additional filtering steps resulted in a final total of 364 hotspot loci having increased likelihood
for a driver or contributory role in hematologic malignancies, which could be confidently used for detecting CHIP.

**Non-cancer cohorts**

Three large cohorts unselected for cancer and having publicly available paired-end whole exome sequencing data were analyzed for CHIP-associated mutations. 1) 1KG: We assessed CHIP from the FASTQ data of 2,535 whole exome sequencing samples coming from the 26 populations of the Phase 3 1000 Genomes project submitted to the European Nucleotide Archive (ENA) as PRJNA262923 by the Wellcome Trust Sanger Institute. Persons in this cohort were generally assumed to be unselected for disease and ≥18 years old. One sample with unavailable paired-end FASTQ data files as well as samples that were poorly sequenced or had other issues previously identified for exclusion (61) were not utilized, resulting in a total of N=2,503 persons whose samples were deemed satisfactory. DNA had been extracted from lymphoblastoid cell lines for most participants (61). Also as per 1000 Genomes Project Consortium, et al. (69), (61), and www.internationalgenome.org, the populations assessed consist of: African Caribbean in Barbados (ACB), Americans of African Ancestry in SW, USA (ASW), Bengali in Bangladesh (BEB), Chinese Dai in Xishuangbanna, China (CDX), Utah Residents (CEPH) with Northern and Western European Ancestry (CEU), Han Chinese in Beijing, China (CHB), Han Chinese South (CHS), Colombians from Medellin, Colombia (CLM), Esan in Nigeria (ESN), Finnish in Finland (FIN), British in England and Scotland (GBR), Gujarati Indian from Houston, Texas (GIH), Gambian in Western Divisions in the Gambia (GWD), Iberian Population in Spain (IBS), Indian Telugu in the UK (ITU), Japanese in Tokyo, Japan (JPT), Kinh in Ho Chi Minh City, Vietnam (KHK), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), Mexican Ancestry from Los Angeles, USA (MXL), Peruvians from Lima, Peru (PEL), Punjabi from...
Lahore, Pakistan (PJL), Puerto Ricans in Puerto Rico (PUR), Sri Lankan Tamil from the UK (STU), Toscani in Italia (TSI), and Yoruba in Ibadan, Nigeria (YRI). 2) SSC: We assessed the FASTQ data of 804 paired-end whole exome sequencing samples from Simons Simplex families deposited in ENA as PRJNA167318. These included 413 parents and 391 children. Of the children, 205 were diagnosed with autism and 186 were unaffected siblings. DNA had been extracted from whole blood as reported in Sanders et al. (60). Of note this cohort has been previously analyzed for postzygotic mutations of substantial frequency (65), and while not specifically looking for CHIP, Lim et al. (65) did identify two of the general (and none of the hotspot) CHIP mutations we list in Supplementary Table S8. 3) QTRG: We assessed paired-end whole exome sequencing FASTQ data from the samples of 1,231 persons of the Qatari genome for precision medicine deposited in ENA as PRJNA290484. These included 510 persons diagnosed with type 2 diabetes, and 270 persons designated as controls. The age spectrum of this cohort ranged from 0-85 years, with 25 persons being <18 years of age. DNA had been extracted from blood as reported in Fakhro et al. (62). While both this QTRG cohort and the SSC cohort assessed DNA from blood, a possibility for infrequent LCL-specific mutations could be present in the data of the 1KG samples. However, we found the frequency of CHIP mutations in the 1KG samples to be similar to the SSC parent cohort, – implying that any LCL-specific effect may be small in magnitude.

CHIP assessment at hotspots

We aligned the paired-end FASTQ data files for each of the 4,538 samples in the non-cancer cohorts to the human reference genome hg19 and performed subsequent deduplication, realignment, recalibration, filtering, and variant calling (additional details provided in Supplementary Data). CHIP at hotspots classification required somatic mutations to have ≥3
variant reads of ≥20 total reads and to occur at one of the 364 confident amino acid hotspots identified in the recurrent mutation analysis (hence at a locus reported in at least three patients and with confident mapping). As the total number of loci at which CHIP could be called was reduced >100-fold from that used in other studies assessing CHIP across large domains (1,092 bases compared to >160,000 bases in Jaiswal et al. (10)), we did not incorporate an additional lower-bound cut-off at hotspots. In addition, such mutations were required to yield the identical amino acid substitution or insertion/deletion effect (e.g. frameshift) as had been previously reported in at least one diagnostic tumor sample, without additional filtering for predicted benign/deleterious effect on protein. Only mutations meeting all of these criteria were used to identify CHIP at hotspots. CHIP at the 70 hotspots with greater potential for artifacts were also assessed, and a recovery analysis was performed to arbitrate any excluded mutations for inclusion. While the coverage statistics at the 364 hotspots were plausibly similar to those of other WES studies analyzed for CHIP, coverage was insufficient to detect small to medium-sized clones at many hotspots. In a separate analysis, a classification of relaxed CHIP at hotspots was used with the sole modification of allowing ≥2 variant reads for identification.

**General CHIP assessment**

We utilized the specified mutations and domains provided in Jaiswal et al. (10) to assess general CHIP (clonal mutations not exclusive to hotspot mutations previously reported in hematologic cancers, though in genes or domains that have been associated with hematopoiesis). We performed filtering for mappability, sequencing artifacts, and recurrence as we did for CHIP at hotspots, with an additional imposed VAF threshold requirement of 4%. We also assessed sensitivity of CHIP prevalence by gene for lower VAF threshold levels. We calculated the
proportion of hotspots in our general CHIP calls at a gene and overall level, and similarly computed these rates on the data of previously published cohorts (5,10).

**Statistical analysis**

We used logistic regression analysis to determine the association of year of age with relaxed CHIP mutation occurrence, with age being mean centered. We also used logistic regression to assess the association of frequency of reported mutations at hotspots and binary detection of CHIP at those loci. As the frequency of reported mutations was right-skewed, the natural logarithm of this variable was used. Sensitivity analyses using unlogged data as well as log base 10 showed similar results. Deviance from expected recurrence frequencies for mutations based on binomial distributions was used to determine the effect of various potential hotspot threshold criteria. Testing for association of variant allele fraction and reported recurrence groups was computed with the Wilcoxon-Mann-Whitney test. A two-sided Fisher’s exact test was used to determine the significance of differences in prevalence for the different variant thresholds and cohort pairings assessed. Additional statistical details are provided in the Supplementary Data. All statistical computations were performed with SAS software, version 9.4 (SAS Institute).
Acknowledgments

We thank the many researchers who published the data and details of their studies, the participants in those studies, and acknowledge their original publications and the European Nucleotide Archive as sources of the primary data. We also thank Allyson Mower and Michele Ballantyne for invaluable advice, and the many publishers who granted text and data mining permission. CCM acknowledges Pediatric Cancer Program funding, – supported by the Intermountain Healthcare and Primary Children’s Hospital Foundations, Department of Pediatrics, and Division of Pediatric Hematology/Oncology at the University of Utah. Some work by JEF was supported by NCATS/NIH (UL1TR002538/TL1TR002540). Some work by SA was funded by a DOD grant (W81XWH-14-1-0417) to BEW. LBJ acknowledges funding from NIH (GM118335/GM059290). The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. The content and expressed viewpoints are those of the authors only.

Authors contributions

JEF, SA, and CCM ran the pipelines. MJB, CVS-W, AF, and CCM performed the literature review and text and data mining. JEF and CCM wrote the first drafts of the manuscript. CCM conceived the study, designed the pipelines, generated the figures, and performed the statistical assessments. JEF, SA, TT, MJB, AF, BEW, JLR-F, CO, LBJ, and JTP provided advice and critically read and edited the manuscript.
References


27. Eisfeld AK, Kohlschmidt J, Schwind S, Nicolet D, Blachly JS, Orwick S, et al. Mutations in the CCND1 and CCND2 genes are frequent events in adult patients with t(8;21)(q22;q22) acute myeloid leukemia. *Leukemia* 2017; **31**: 1278-1285.


Table 1. The 48 studies included in the hotspot mutation analysis to identify recurrent mutations in hematologic malignancies. Studies assessed somatic mutations in diagnostic patient samples with next generation sequencing and met additional inclusion criteria.

<table>
<thead>
<tr>
<th>Study (year)</th>
<th>Main hematologic malignancy</th>
<th>No. patients assessed</th>
<th>No. mutations assessed</th>
<th>No. protein-altering mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andersson et al (2015)¹¹</td>
<td>ALL</td>
<td>65</td>
<td>250</td>
<td>146</td>
</tr>
<tr>
<td>Chen et al (2018)⁴⁵</td>
<td>ALL</td>
<td>36</td>
<td>302</td>
<td>294</td>
</tr>
<tr>
<td>De Keersmaecker et al (2015)¹⁵</td>
<td>ALL</td>
<td>211</td>
<td>538</td>
<td>470</td>
</tr>
<tr>
<td>Holmfeldt et al (2013)¹⁴</td>
<td>ALL</td>
<td>40</td>
<td>706</td>
<td>366</td>
</tr>
<tr>
<td>Liu et al (2016)¹³</td>
<td>ALL</td>
<td>203</td>
<td>2,437</td>
<td>1,418</td>
</tr>
<tr>
<td>Liu et al (2017)¹⁶</td>
<td>ALL</td>
<td>264</td>
<td>4,165</td>
<td>3,426</td>
</tr>
<tr>
<td>Oshima et al (2016)¹⁰</td>
<td>ALL</td>
<td>55</td>
<td>1,845</td>
<td>397</td>
</tr>
<tr>
<td>Papaemmanuil et al (2014)¹⁶</td>
<td>ALL</td>
<td>55</td>
<td>795</td>
<td>549</td>
</tr>
<tr>
<td>Paulsson et al (2015)¹⁹</td>
<td>ALL</td>
<td>51</td>
<td>459</td>
<td>376</td>
</tr>
<tr>
<td>Russell et al (2017)²⁰</td>
<td>ALL</td>
<td>8</td>
<td>218</td>
<td>86</td>
</tr>
<tr>
<td>Ryan et al (2016)²²</td>
<td>ALL</td>
<td>42</td>
<td>45</td>
<td>30</td>
</tr>
<tr>
<td>Stengel et al (2014)²³</td>
<td>ALL</td>
<td>625</td>
<td>110</td>
<td>98</td>
</tr>
<tr>
<td>Ma et al (2018)³</td>
<td>ALL</td>
<td>122</td>
<td>574</td>
<td>254</td>
</tr>
<tr>
<td>Huether et al (2014)²⁴</td>
<td>ALL</td>
<td>525</td>
<td>412</td>
<td>167</td>
</tr>
<tr>
<td>Bolouri et al (2018)²⁵</td>
<td>AML</td>
<td>684</td>
<td>1,983</td>
<td>584</td>
</tr>
<tr>
<td>de Rood et al (2017)²⁶</td>
<td>AML</td>
<td>113</td>
<td>268</td>
<td>99</td>
</tr>
<tr>
<td>Eisfeld et al (2017)²³</td>
<td>AML</td>
<td>177</td>
<td>258</td>
<td>199</td>
</tr>
<tr>
<td>Eisfeld et al (2017)²⁴</td>
<td>AML</td>
<td>10</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Eisfeld et al (2016)²⁵</td>
<td>AML</td>
<td>23</td>
<td>68</td>
<td>56</td>
</tr>
<tr>
<td>Faber et al (2016)²⁶</td>
<td>AML</td>
<td>165</td>
<td>849</td>
<td>588</td>
</tr>
<tr>
<td>Farrar et al (2016)²¹</td>
<td>AML</td>
<td>20</td>
<td>208</td>
<td>126</td>
</tr>
<tr>
<td>Greif et al (2018)²⁴</td>
<td>AML</td>
<td>50</td>
<td>558</td>
<td>440</td>
</tr>
<tr>
<td>Hirsch et al (2016)²⁵</td>
<td>AML</td>
<td>53</td>
<td>257</td>
<td>246</td>
</tr>
<tr>
<td>CGARN et al (2013)³⁶</td>
<td>AML</td>
<td>200</td>
<td>22,633</td>
<td>1,574</td>
</tr>
<tr>
<td>Madan et al (2016)¹²</td>
<td>AML</td>
<td>153</td>
<td>210</td>
<td>126</td>
</tr>
<tr>
<td>Sehgal et al (2015)²⁵</td>
<td>AML</td>
<td>152</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Sood et al (2016)²⁶</td>
<td>AML</td>
<td>13</td>
<td>180</td>
<td>94</td>
</tr>
<tr>
<td>Thol et al (2017)¹⁷</td>
<td>AML</td>
<td>171</td>
<td>603</td>
<td>415</td>
</tr>
<tr>
<td>Kim et al (2017)²⁷</td>
<td>CML</td>
<td>100</td>
<td>68</td>
<td>40</td>
</tr>
<tr>
<td>Togasaki et al (2017)²⁸</td>
<td>CML</td>
<td>24</td>
<td>191</td>
<td>181</td>
</tr>
<tr>
<td>Mose et al (2016)²⁹</td>
<td>CML</td>
<td>69</td>
<td>507</td>
<td>479</td>
</tr>
<tr>
<td>Merlevede et al (2016)²⁹</td>
<td>CMLL</td>
<td>17</td>
<td>8,077</td>
<td>143</td>
</tr>
<tr>
<td>Palomo et al (2016)²⁶</td>
<td>CMLL</td>
<td>56</td>
<td>262</td>
<td>255</td>
</tr>
<tr>
<td>Patnaik et al (2017)²⁷</td>
<td>CMLL</td>
<td>261</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Haferlach et al (2014)²⁰</td>
<td>MDS</td>
<td>102</td>
<td>300</td>
<td>281</td>
</tr>
<tr>
<td>Pastor et al (2017)²⁸</td>
<td>MDS</td>
<td>50</td>
<td>25</td>
<td>18</td>
</tr>
<tr>
<td>Walter et al (2013)²⁹</td>
<td>MDS</td>
<td>150</td>
<td>322</td>
<td>254</td>
</tr>
<tr>
<td>Yoshida et al (2011)²¹</td>
<td>MDS</td>
<td>22</td>
<td>268</td>
<td>185</td>
</tr>
<tr>
<td>Churpek et al (2015)²⁴</td>
<td>MDS</td>
<td>16</td>
<td>183</td>
<td>51</td>
</tr>
<tr>
<td>Lundberg et al (2014)²⁶</td>
<td>MFPN</td>
<td>197</td>
<td>267</td>
<td>236</td>
</tr>
<tr>
<td>Nangalia et al (2013)²³</td>
<td>MFPN</td>
<td>151</td>
<td>1,498</td>
<td>1,064</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>7,430</td>
<td>58,177</td>
<td>20,141</td>
</tr>
</tbody>
</table>
*Number of unique diagnostic patients assessed in the original study meeting inclusion criteria (hence, the total sample size of the original study may have been larger). †Number of mutations from the original study assessed in this hotspot mutation analysis. ‡Number of determined protein-altering or splice site mutations used in recurrence tallies. §CGARN = Cancer Genome Atlas Research Network. ¶Sample size estimated for the random selection of mutations provided. §Some samples from patients classified as CMML were included in this study. **Some patients were also classified with AML in this study. ***Some patients were also classified with AML or MPN/JMML in this study.
Figure legends

**Fig. 1. Flowchart of mutational hotspot and CHIP analyses.** Hematologic cancer hotspot mutations identified in the recurrent mutation analysis were assessed for clonal presence in persons from non-cancer cohorts as part of the CHIP analysis.

**Fig. 2. Mutation hotspots in hematologic malignancies.** (A) Most frequently reported mutation hotspots at blood cancer diagnoses. Top axes provide the total patients assessed for each hematologic malignancy. Numbers within cells: total number of patients having a reported mutation. Color intensity scale: percentage of persons having a reported mutation for each malignancy. As not all patients were assessed for mutations at each locus, these frequencies likely represent lower-bound estimates. Yellow boxes indicate mutations identified in only a single hematologic malignancy. (B) Location in coding region for the 434 hotspots occurring in 85 genes. Horizontal axes: normalized location of each mutation locus within the coding sequence. Color of triangles: types of mutations reported at each locus; purple indicates nonsense or frameshift only, blue indicates only point mutations or non-frameshift, orange indicates a mixture of nonsense/frameshift and point/non-frameshift, pink indicates splice site.

**Fig. 3. CHIP detection at hotspots in non-cancer cohorts.** (A) Proportion of each cohort with identified CHIP mutations identical to mutations reported at 364 confident mutation hotspots (black colored bars indicate cohort aggregates; for sub-cohort abbreviations, see Materials and Methods). (B) Genes with CHIP mutations identical to mutations reported at hotspots across the three cohorts (N=4,530). Black color indicates mutation identified in a child sample. (C) Rate of relaxed CHIP (mutations having ≥2 variant reads) in the QTRG cohort by age-groups.
Recruent Mutation (Hotspot) Analysis

Initial extraction evaluation
14 studies

PubMed search for NGS mutation studies of blood cancers
172 studies

Single transcript determination

Studies evaluated for providing:
• NGS assessment of diagnostic pt. sample
• unambiguous genetic coordinates
• consistent genetic mapping

48 useable studies

All mutations harmonized to single transcript;
identified protein altering and splice site
mutations with locus reported in three or
more patients
434 mutation hotspots

364 hotspots identifiable for CHIP (confident loci)

70 hotspots with high potential for artifactual CHIP calls

Clonal Hematopoiesis of Indeterminate Potential (CHIP) Analysis

1KG cohort (non-cancer)
2503 persons

Simons Simplex Collection (non-cancer)
804 persons

Qatari cohort (non-cancer)
1,231 persons

4,538 persons with paired-end whole exome sequencing data (8 persons later excluded)

Data processed to identify CHIP at hotspots

Data processed to identify CHIP not exclusive to hotspots (as per ref. 10)

CHIP mutations identified in:
• 83 persons with CHIP at hotspots (82 at confident loci, 1 in recovery analysis of high potential for artifact loci)
• 183 persons with CHIP not exclusive to hotspots (6 persons with 2 CHIP mutations)

364 hotspots identifiable for CHIP (confident loci)
Large-Scale Identification of Clonal Hematopoiesis and Mutations Recurrent in Blood Cancers

Julie E Feusier, Sasi Arunachalam, Tsewang Tashi, et al.

Blood Cancer Discov Published OnlineFirst March 3, 2021.

Updated version
Access the most recent version of this article at:
doi: 10.1158/2643-3230.BCD-20-0094

Supplementary Material
Access the most recent supplemental material at:
http://bloodcancerdiscov.aacrjournals.org/content/suppl/2021/02/26/2643-3230.BCD-20-0094.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link http://bloodcancerdiscov.aacrjournals.org/content/early/2021/02/25/2643-3230.BCD-20-0094. Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.