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 usable for CHIP detection by performing a data mining analysis of 48 somatic mutation landscape studies reporting mutations at diagnoses of 7,430 adult and pediatric patients with leukemia or other hematologic malignancy. 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 noncancer 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 precancer surveillance for both children and adults.

SIGNIFICANCE: This study identifies frequently occurring mutations across several blood cancers that 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 noncancer cohorts, including children, showcasing potential for improved precancer surveillance.

See related commentary by Spitzer and Levine.
INTRODUCTION

Somatic mutations at hotspots (genetic loci observed to be frequently mutated across patients with cancer) 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 hematologic 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), https://cancer.sanger.ac.uk/cosmic; ref. 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 malignancy 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 (HR >10) for future blood cancer development (5, 6). CHIP increases sharply in prevalence with advanced age (5–9), and backward 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 (MPN), 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 noncancer cohorts, which included the widely used 1000 Genomes Project (1KG) and more than 400 children. Our findings may lead to the improved prognostic and clinical utility of precancer 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 that may have been preceded by CHIP (58), we limited our investigation to studies that 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 (Methods; 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 multistudy 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 patients with AML, 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 observed only in patients with ALL. We also observed
### Table 1. The 48 studies included in the hotspot mutation analysis to identify recurrent mutations in hematologic malignancies

<table>
<thead>
<tr>
<th>Study</th>
<th>Main hematologic malignancy</th>
<th>Patients assessed (N)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Mutations assessed (N)&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Protein-altering mutations (N)&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andersson et al. (2015; ref. 11)</td>
<td>ALL</td>
<td>65</td>
<td>250</td>
<td>146</td>
</tr>
<tr>
<td>Chen et al. (2018; ref. 12)</td>
<td>ALL</td>
<td>36</td>
<td>302</td>
<td>294</td>
</tr>
<tr>
<td>De Keersmaecker et al. (2013; ref. 13)</td>
<td>ALL</td>
<td>211</td>
<td>538</td>
<td>470</td>
</tr>
<tr>
<td>Holmfeldt et al. (2013; ref. 14)</td>
<td>ALL</td>
<td>40</td>
<td>706</td>
<td>366</td>
</tr>
<tr>
<td>Liu et al. (2016; ref. 15)</td>
<td>ALL</td>
<td>203</td>
<td>2,437</td>
<td>1,418</td>
</tr>
<tr>
<td>Liu et al. (2017; ref. 16)</td>
<td>ALL</td>
<td>264</td>
<td>4,165</td>
<td>3,426</td>
</tr>
<tr>
<td>Oshima et al. (2016; ref. 17)</td>
<td>ALL</td>
<td>55</td>
<td>1,845</td>
<td>397</td>
</tr>
<tr>
<td>Papaemmanuil et al. (2014; ref. 18)</td>
<td>ALL</td>
<td>55</td>
<td>795</td>
<td>549</td>
</tr>
<tr>
<td>Paulsson et al. (2015; ref. 19)</td>
<td>ALL</td>
<td>51</td>
<td>459</td>
<td>376</td>
</tr>
<tr>
<td>Russell et al. (2017; ref. 20)</td>
<td>ALL</td>
<td>8</td>
<td>218</td>
<td>86</td>
</tr>
<tr>
<td>Ryan et al. (2016; ref. 21)</td>
<td>ALL</td>
<td>42</td>
<td>45</td>
<td>30</td>
</tr>
<tr>
<td>Stengel et al. (2014; ref. 22)</td>
<td>ALL</td>
<td>625</td>
<td>110</td>
<td>98</td>
</tr>
<tr>
<td>Ma et al. (2018; ref. 3)</td>
<td>ALL</td>
<td>122</td>
<td>574</td>
<td>254</td>
</tr>
<tr>
<td>Huether et al. (2014; ref. 23)</td>
<td>ALL</td>
<td>525</td>
<td>412</td>
<td>167</td>
</tr>
<tr>
<td>Bolouri et al. (2018; ref. 24)</td>
<td>AML</td>
<td>684</td>
<td>1,983</td>
<td>584</td>
</tr>
<tr>
<td>de Rooij et al. (2017; ref. 25)</td>
<td>AML</td>
<td>113</td>
<td>268</td>
<td>99</td>
</tr>
<tr>
<td>Dolnik et al. (2012; ref. 26)</td>
<td>AML</td>
<td>50</td>
<td>171</td>
<td>153</td>
</tr>
<tr>
<td>Eisfeld et al. (2017; ref. 27)</td>
<td>AML</td>
<td>177</td>
<td>258</td>
<td>199</td>
</tr>
<tr>
<td>Eisfeld et al. (2017; ref. 28)</td>
<td>AML</td>
<td>10</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Eisfeld et al. (2016; ref. 29)</td>
<td>AML</td>
<td>23</td>
<td>68</td>
<td>56</td>
</tr>
<tr>
<td>Faber et al. (2016; ref. 30)</td>
<td>AML</td>
<td>165</td>
<td>849</td>
<td>588</td>
</tr>
<tr>
<td>Farrar et al. (2016; ref. 31)</td>
<td>AML</td>
<td>20</td>
<td>208</td>
<td>126</td>
</tr>
<tr>
<td>Garg et al. (2015; ref. 32)</td>
<td>AML</td>
<td>67</td>
<td>443</td>
<td>250</td>
</tr>
<tr>
<td>Greif et al. (2018; ref. 33)</td>
<td>AML</td>
<td>50</td>
<td>558</td>
<td>440</td>
</tr>
<tr>
<td>Hirsch et al. (2016; ref. 34)</td>
<td>AML</td>
<td>53</td>
<td>257</td>
<td>246</td>
</tr>
<tr>
<td>Lavallée et al. (2015; ref. 35)</td>
<td>AML</td>
<td>29</td>
<td>55</td>
<td>53</td>
</tr>
<tr>
<td>CGARN et al. (2013; ref. 36)</td>
<td>AML</td>
<td>200</td>
<td>22,633</td>
<td>1,574</td>
</tr>
<tr>
<td>Madan et al. (2016; ref. 37)</td>
<td>AML</td>
<td>153</td>
<td>210</td>
<td>126</td>
</tr>
<tr>
<td>Papaemmanuil et al. (2016; ref. 38)</td>
<td>AML</td>
<td>1,540</td>
<td>3,902</td>
<td>3,341</td>
</tr>
<tr>
<td>Sehgal et al. (2015; ref. 39)</td>
<td>AML</td>
<td>152</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Sood et al. (2016; ref. 40)</td>
<td>AML</td>
<td>13</td>
<td>180</td>
<td>94</td>
</tr>
<tr>
<td>Thol et al. (2017; ref. 41)</td>
<td>AML</td>
<td>171</td>
<td>603</td>
<td>415</td>
</tr>
<tr>
<td>Kim et al. (2017; ref. 42)</td>
<td>CML</td>
<td>100</td>
<td>68</td>
<td>40</td>
</tr>
<tr>
<td>Togasaki et al. (2017; ref. 43)</td>
<td>CML</td>
<td>24</td>
<td>191</td>
<td>181</td>
</tr>
<tr>
<td>Mason et al. (2016; ref. 44)</td>
<td>CMML</td>
<td>69</td>
<td>507</td>
<td>479</td>
</tr>
<tr>
<td>Merlevede et al. (2016; ref. 45)</td>
<td>CMML</td>
<td>17</td>
<td>8,077</td>
<td>143</td>
</tr>
<tr>
<td>Palomo et al. (2016; ref. 46)</td>
<td>CMML</td>
<td>56</td>
<td>262</td>
<td>255</td>
</tr>
<tr>
<td>Patnaik et al. (2017; ref. 47)</td>
<td>CMML</td>
<td>261</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Caye et al. (2015; ref. 48)</td>
<td>JMML</td>
<td>118</td>
<td>187</td>
<td>107</td>
</tr>
<tr>
<td>Stiegitz et al. (2015; ref. 49)</td>
<td>JMML</td>
<td>71</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>Haferlach et al. (2014; ref. 50)</td>
<td>MDS</td>
<td>102&lt;sup&gt;a&lt;/sup&gt;</td>
<td>300</td>
<td>281</td>
</tr>
<tr>
<td>Pastor et al. (2017; ref. 51)</td>
<td>MDS</td>
<td>50</td>
<td>25</td>
<td>18</td>
</tr>
</tbody>
</table>

(continued)
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 sizable 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 23 patients at diagnosis of a hematologic malignancy (Supplementary Table S5). Many of these unique mutations occur within NOTCH1, FLT3, CEBPA, KIT, and RUNX1.

CHIP in Noncancer Cohorts Identical to Hotspot Mutations

We next assessed three large noncancer cohorts (60–62) for CHIP at our identified hotspots within their combined 4,538 individuals. The 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 eight 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 as follows: 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 [OR, 1.75 per log of reported mutations; 95% confidence interval (CI), 1.33–2.29; 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 noncancer
Clonal Hematopoiesis and Recurrent Blood Cancer Mutations

PubMed search for NGS mutation studies of blood cancers
172 studies

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

Coasts, 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 of 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 RUNX1 p.S322X clonal mutation also had an apparent germline SNP at RUNX1 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 to 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

Figure 1. Flowchart of mutational hotspot and clonal hematopoiesis of indeterminate potential (CHIP) analyses. Hematologic cancer hotspot mutations identified in the recurrent mutation analysis were assessed for clonal presence in persons from noncancer cohorts as part of the CHIP analysis. pt., patient.
Figure 2. Mutation hotspots in hematologic malignancies. A, Most frequently reported mutation hotspots at blood cancer diagnoses. Top axes: 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 only nonsense or frameshift, blue indicates only point mutations or nonframeshift, orange indicates a mixture of nonsense/frameshift and point/nonframeshift, and pink indicates splice site). Utilized transcripts are provided in Supplementary Table S1.
Clonal Hematopoiesis and Recurrent Blood Cancer Mutations

RCH Brief

Clonal Hematopoiesis and Recurrent Blood Cancer Mutations (37.0%) reported CHIP mutations in Jaiswal and colleagues listed in Supplementary Table S4). For comparison, we also identified hotspots in three or more diagnostic patients in the analyzed cohorts (Supplementary Table S8), 40 of which were not exclusive to hotspots within the 4,530 persons of our cohort, and, most strikingly, DNMT3A, TET2, ASXL1 and KMT2A not exclusively in hotspots in our hotspot list. As in previous studies, CHIP not exclusive to hotspots was most frequent for TET2, ASXL1 and JAK2 mutations were noticeably fewer underlying cases of cardiovascular disease and younger previously observed (10). As the cohorts we assessed likely had between control and cardiovascular disease cohorts was pre-

Our general CHIP analysis identified six persons having two mutations (five of whom had a mutation of 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 and colleagues (10) 193/4,530 (42.6%) persons had qualifying CHIP. As in previous studies, CHIP not exclusive to hotspots was most frequent for TET2 and DNMT3A (Supplementary Table S8). However, we found mutations in SETD2, EP300, and KMT2A/D to be more common in our cohort, whereas ASXL1 and JAK2 mutations were noticeably less frequent. Variability of gene prevalence for CHIP mutations in DNMT3A, TET2, ASXL1, and, most strikingly, 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

associated with relaxed CHIP prevalence in this call set (OR, 1.025 per year of age; 95% CI, 1.004–1.048; P = 0.022; Fig. 3C).

CHIP Not Exclusive to Hotspots

We also investigated the noncancer 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 and colleagues (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 three 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 (37.0%) reported CHIP mutations in Jaiswal and colleagues (5) and Jaiswal and colleagues (10), respectively, were identical to mutations in our hotspot list.

Figure 3. CHIP detection at hotspots in noncancer 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). ACR, African Caribbean in Barbados; ASW, Americas of African Ancestry in SW, USA; BEB, Bengali in Bangladesh; CDX, Chinese Dai in Xishuangbanna, China; CEU, Utah Residents with Northern and Western European Ancestry; CHB, Han Chinese in Beijing, China; CHS, Han Chinese South; CLM, Colombians from Medellin, Colombia; ESN, Esan in Nigeria; FIN, Finnish in Finland; GBR, British in England and Scotland; GJH, Gujarati Indian from Houston, Texas; GWD, Gambian in Western Divisions in the Gambia; IBS, Iberian Population in Spain; ITU, Indian Telugu in the UK; JPT, Japanese in Tokyo, Japan; KHV, Kinh in Ho Chi Minh City, Vietnam; UWE, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; MXL, Mexican Ancestry from Los Angeles, USA; PEL, Peruvians from Lima, Peru; PPL, Punjabi from Lahore, Pakistan; PUR, Puerto Ricans in Puerto Rico; STU, Sri Lankan Tamil from the UK; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria. 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 Qatari cohort by age groups.
average age than other CHIP studies using WES data, some of our gene prevalence rates may reflect these differences. Further exploration identified seven ASXL1 variants having a VAF >2%, yet only one reached the threshold of 4% required for this general CHIP analysis. No additional 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 (DNMT3A p.V665L) had been previously identified and reported as a de novo mutation by Lim and colleagues (65), although it could possibly have been a CHIP mutation that had reached near-complete saturation (the VAF was 47.3%; see 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 nonhotspot 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 hotspot regions we identified identified to have a hotspot CHIP mutation, making the child’s nonhotspot 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.

**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 usable for hotspot determination. This preliminary work was supplemented with a formal PubMed search on studies published before July 2018 with the following search criteria: (“genomic landscape,” “somatic landscape,” “genomic profile,” “mutational landscape,” OR...
Clonal Hematopoiesis and Recurrent Blood Cancer Mutations

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 reanalysis. 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 VAFs 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 criteria deemed acceptable were 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 reanalysis. 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.

Noncancer Cohorts

Three large cohorts unscreened for cancer and having publicly available paired-end WES data were analyzed for CHIP-associated mutations. (i) 1KG: We assessed CHIP from the FASTQ data of 2,535 WES 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 (LCL) for most participants (61). Also as per 1000 Genomes Project Consortium and colleagues (61, 69) and The International Genome Sample Resource (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 (from CEPH families) with Northern and Western European Ancestry (CEU); Han Chinese in Beijing, China (CHB); Han Chinese South (CHS); Colombians from Medelin, Colombia (CLM); 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 (KHV); 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 (PIL); Puerto Ricans in Puerto Rico (PUR); Sri Lankan Tamil from the UK (STU); Toscani in Italia (TSI); and Yoruba in Ibadan, Nigeria (YRI). (ii) SSC: We assessed the FASTQ data of 804 paired-end WES 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 and colleagues (60). Of note, this cohort has been previously analyzed for postzygotic mutations of substantial frequency (65), and while not specifically looking for CHIP, Lim and colleagues (65) did attempt to identify two of the general (and none of the hotspot) CHIP mutations we list in Supplementary Table S8. (iii) QTGR: We assessed paired-end WES FASTQ data from the samples of 1,231
persons of the QTRG for precision medicine deposited in ENA as PRJNA290484. These included 510 persons diagnosed with type II diabetes and 270 persons designated as controls. The age spectrum of this cohort ranged from 0 to 85 years, with 25 persons being <18 years of age. DNA had been extracted from blood as reported in Fakhro and colleagues (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 noncancer cohorts to the human reference genome hg19 and performed subsequent deduplication, realignment, recalibration, filtering, and variant calling (additional details provided in Supplementary Methods). CHIP at hotspots classification required somatic mutations to have ≥23 variant reads of ≥220 total reads and to occur at one of the 364 confident amino acid hotspot 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 with >160,000 bases in Jaiswal and colleagues; ref. 10), we did not incorporate an additional lower-bound cutoff 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 and colleagues (10) to assess general CHIP (clonal mutations not exclusive to hotspot mutations previously reported in hematologic cancers, although 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 level 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 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 Methods. All statistical computations were performed with SAS software, version 9.4 (SAS Institute).

**Authors’ Disclosures**

J.L. Rodriguez-Flores is a full-time employee of Regeneron Pharmaceuticals Inc. C.C. Mason reports other from Intermountain Healthcare Foundation (funding for the Pediatric Cancer Program) and Primary Children’s Hospital Foundation (funding for the Pediatric Cancer Program), and grants from Primary Children’s Center for Personalized Medicine during the conduct of the study. No disclosures were reported by the other authors.

**Disclaimer**

The content and expressed viewpoints are those of the authors only.

**Authors’ Contributions**

J.E. Feusier: Data curation, software, validation, investigation, writing—original draft, writing—review and editing. S. Arunachalam: Data curation, software, validation, investigation, writing—review and editing. T. Tashi: Conceptualization, writing—review and editing. M.J. Baker: Data curation, writing—review and editing. C. VanSant-Webb: Data curation. A. Ferdig: Data curation, writing—review and editing. B.E. Welm: Funding acquisition, writing—review and editing. L.B. Jorde: Conceptualization, funding acquisition, writing—review and editing. J.T. Prechal: Conceptualization, writing—review and editing. C.C. Mason: Conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing—original draft, project administration, writing—review and editing.

**Acknowledgments**

The authors thank the many researchers who published the data and details of their studies as well as the participants in those studies, and acknowledge their original publications and the European Nucleotide Archive as sources of the primary data. They also thank Allyson Mower and Michele Ballantyne for invaluable advice, and the many publishers who granted text and data mining permission. C.C. Mason acknowledges Pediatric Cancer Program funding, which is supported by the Intermountain Healthcare and Primary Children’s Hospital Foundations as well as the Department of Pediatrics and Division of Pediatric Hematology/Oncology at the University of Utah. Some work by J.E. Feusier was supported by National Center for Advancing Translational Sciences (NCATS)/NIH (UL1TR002538/TLTR002540). Some work by S. Arunachalam was funded by a U.S. Department of Defense grant (W81XWH-14-1-0417; to B.E. Welm). L.B. Jorde 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.

Received June 5, 2020; revised December 21, 2020; accepted February 24, 2021; published first March 3, 2021.

**REFERENCES**

Clonal Hematopoiesis and Recurrent Blood Cancer Mutations

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.