



Published in final edited form as:

Nat Med. 2014 December ; 20(12): 1472–1478. doi:10.1038/nm.3733.

## Age-related cancer mutations associated with clonal hematopoietic expansion

Mingchao Xie<sup>1,2,\*</sup>, Charles Lu<sup>1,\*</sup>, Jiayin Wang<sup>1,2,\*</sup>, Michael D. McLellan<sup>1</sup>, Kimberly J. Johnson<sup>3</sup>, Michael C. Wendl<sup>1,4,5</sup>, Joshua F. McMichael<sup>1</sup>, Heather K. Schmidt<sup>1</sup>, Venkata Yellapantula<sup>1,2</sup>, Christopher A. Miller<sup>1</sup>, Bradley A. Ozenberger<sup>1,2</sup>, John S. Welch<sup>2,6</sup>, Daniel C. Link<sup>2,6</sup>, Matthew J. Walter<sup>2,6</sup>, Elaine R. Mardis<sup>1,2,4,6</sup>, John F. Dpersio<sup>2,6</sup>, Feng Chen<sup>2,6</sup>, Richard K. Wilson<sup>1,2,4,6</sup>, Timothy J. Ley<sup>1,2,4,6</sup>, and Li Ding<sup>1,2,4,6,#</sup>

<sup>1</sup>The Genome Institute, Washington University in St. Louis, MO 63108, USA

<sup>2</sup>Department of Medicine, Washington University in St. Louis, MO 63108, USA

<sup>3</sup>Brown School Master of Public Health Program, Washington University in St. Louis, St. Louis, MO 63130, USA

<sup>4</sup>Department of Genetics, Washington University in St. Louis, MO 63108, USA

<sup>5</sup>Department of Mathematics, Washington University in St. Louis, MO 63108, USA

<sup>6</sup>Siteman Cancer Center, Washington University in St. Louis, MO 63108, USA

### Abstract

Several genetic alterations characteristic of leukemia and lymphoma have been detected in the blood of individuals without apparent hematological malignancies. We analyzed blood-derived sequence data from 2,728 individuals within The Cancer Genome Atlas, and discovered 77 blood-specific mutations in cancer-associated genes, the majority being associated with advanced age. Remarkably, 83% of these mutations were from 19 leukemia/lymphoma-associated genes, and nine were recurrently mutated (*DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *TP53*, *GNAS*, *PPM1D*, *BCORL1* and *SF3B1*). We identified 14 additional mutations in a very small fraction of blood cells, possibly representing the earliest stages of clonal expansion in hematopoietic stem cells. Comparison of these findings to mutations in hematological malignancies identified several recurrently mutated genes that may be disease initiators. Our analyses show that the blood cells of more than 2% of individuals (5–6% of people older than 70 years) contain mutations that may represent premalignant, initiating events that cause clonal hematopoietic expansion.

---

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:[http://www.nature.com/authors/editorial\\_policies/license.html#terms](http://www.nature.com/authors/editorial_policies/license.html#terms)

#Corresponding Author: Li Ding, Ph.D., The Genome Institute, Division of Oncology, Department of Medicine, Washington University School of Medicine, St. Louis, MO 63108, [lding@genome.wustl.edu](mailto:lding@genome.wustl.edu).

\*These authors contributed equally to this work.

### CONTRIBUTIONS

L.D. designed and supervised research. M.X., C.L., J.W., M.D.M., M.C.W., H.K.S., V.Y., C.A.M., J.S.W., D.C.L., M.J.W., T.J.L., F.C., K.J.J., and L.D. analyzed the data. T.J.L., J.S.W., D.C.L., M.J.W., and J.F.D. provided disease specific analysis. M.C.W. performed statistical analysis. E.R.M. and R.K.W. directed sequencing experiments. M.X., M.D.M., C.L., M.C.W., and J.F.M. prepared figures and tables. L.D. and F.C. wrote the manuscript. L.D., T.J.L., and M.C.W. revised the manuscript.

## INTRODUCTION

Blood cells are continuously regenerated by hematopoietic stem/progenitor cells (HSPCs). Human HSPCs divide only rarely (estimated at once a month), but have self-renewal properties that sustain survival for decades. As HSPCs divide, they accumulate rare, random mutations that generally do not affect function<sup>1</sup>. However, some mutations confer advantages in self-renewal and/or proliferation, resulting in clonal expansion of the affected cells. Although these “initiating” mutations do not lead directly to disease, they can cooperate with subsequent mutations to cause hematopoietic malignancies. For example, *BCR-ABL* and *BCL2* translocations have been found in blood cells of individuals without overt hematological malignancies<sup>2–4</sup>. The frequency of such events appears to increase with age, with a similar trend being found for somatic structural changes in the nuclear genomes of blood cells<sup>1,5</sup>. SNP array analysis from large GWAS cohorts showed ~2–3% of normal individuals of advanced age (70s and 80s) harbor leukemia-associated copy number changes that include genes such as *DNMT3A* and *TET2*<sup>6,7</sup>. More recently, somatic recurrent *TET2* mutations were detected in the blood of elderly women without overt hematological malignancies<sup>8</sup> and *DNMT3A* mutation was reported in non-leukemic cells<sup>9</sup>.

These findings have collectively led to the hypothesis that certain genetic mutations may confer advantages to affected HSPCs, resulting in enhanced cell renewal and/or clonal expansion. However, it is unclear whether the effect involves only a small number of genes, or many more genes related to leukemia/lymphoma, and whether their participation in promoting clonal expansion necessarily leads to clones resembling cancer cells. While analyzing variations in 2,728 TCGA blood samples, we observed many individuals with age-related hematopoietic clonal mosaicism and concurrent presence of over 60 mutations in 19 leukemia/lymphoma-associated genes. Our study identified not only genes, but also specific mutations associated with the clonal expansion process. Additional statistical analysis identified low-level (2 to 10% variant allele fractions) recurrent leukemic mutations in a substantial number of cases, possibly in the early stages of clonal expansion. Moreover, our analysis suggests that *DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *SF3B1*, and *TP53* have distinct and overlapping roles in the development of MPN, MDS, CLL, and/or AML. Finally, these results also incidentally highlight the need for caution when using blood as a reference for a surrogate “germline” genome, especially in older individuals.

## RESULTS

### Cancer types and sample characteristics

We searched for variants present in the blood normal controls across 2,728 cancer patients (Supplementary Table 1a) from 11 diverse cancer types: breast adenocarcinoma (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), brain low grade glioma (LGG), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian carcinoma (OV), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD), and uterine corpus endometrioid carcinoma (UCEC). The numbers of cases from each tumor type range from 57 (KIRC) to 673 (BRCA) and are listed in Supplementary Table 1b. Patients were diagnosed between 10–90 years (mean  $59.5 \pm 13.1$  years) and 22.1% were deceased at the

time of TCGA sample procurement (Supplementary Table 1b). TCGA collects clinical data regarding diagnosis and prior treatment of neoplasms during the sample submission process. To ensure that our dataset was comprised only of individuals having *first-time* primary cancers and having had no treatment with radiation and/or chemotherapy, we excluded those having reported histories of these events as identified at <https://tcga-data.nci.nih.gov/annotations/> and all clinical data (July 30<sup>th</sup>, 2014). However, five patients with synchronous tumors not associated with blood were included, since these synchronous tumors would be unlikely to affect variant analysis in corresponding blood samples.

### Variant calling and filtering strategies

Variants in the 2,728 blood normal controls were identified with VarScan (single nucleotide variant and indel), GATK (single nucleotide variant and indel), and Pindel (indel) (see Methods). False-positive filters were subsequently applied prior to downstream analysis and interpretation (see Methods). Out of the 49,317,027 variants (previously reported OV counts<sup>10</sup> were not included here) that passed false positive filters, 1,622,485 with minor allele frequency of <1% in the 1000 Genomes reference and in each cancer cohort were retained for further analysis; this consists of 1,025,632 missense, 529,505 synonymous, 19,663 nonsense, 10,976 splice site, 926 nonstop/readthrough, 20,275 frameshift indels, and 15,508 in frame indels (Supplementary Table 1c). We used a stringent filtering strategy described previously<sup>11</sup> for standardizing specificity across the Pan-Cancer somatic variant calls for available matched tumor samples (Supplementary Table 2).

### Variants contributing to hematopoietic clonal expansion

The collection of both tumor and matched blood normal exome data by TCGA provides a unique comparative resource for identifying those somatic variants in blood that contribute to clonal expansion. We set out to identify both rare truncation variants (RTV), i.e. those having <1% MAF in both the 1000 Genomes collection and the cohort data, and variants overlapping with recurrent somatic mutations (also called Known Hotspot Variants: KHV) found in the analysis of 12 TCGA cancer types (see Methods and Supplementary Table 3a and 3b). Subsequently, 1,598 RTVs in 556 cancer-associated genes based on several recent studies<sup>11–15</sup> (see Methods and Supplementary Table 4) were manually reviewed to remove false positive calls and 136 KHVs in the same set of cancer genes were identified. The resulting list was further filtered to remove polymorphisms present in 1000 Genomes (see Methods) and having greater than 0.1% MAF as reported in the current Exome Variant Server data release (ESP6500SI-V2, <http://evs.gs.washington.edu/EVS/>) from 6,503 samples drawn from multiple NHLBI Exome Sequencing Project (ESP) cohorts. We focused on those mutations found in blood normal samples, but not present or present only at very low levels in either the tumor samples or tumor adjacent normal samples, as this pattern is highly suggestive of somatic mutation in HSPCs introduced by the clonal expansion process. Inflammatory lymphocytes/macrophages/neutrophils will infiltrate different tumors to different extents. Therefore, the hematopoietic mutations do not have to be completely absent in the tumor sample.

Our analysis of RTVs and KHVs in 556 selected cancer-associated genes identified 70 blood-specific mutations in 58 individuals. We further performed comparative analysis of

blood versus tumor samples for these 58 individuals, with the goal of detecting all blood-specific nonsynonymous mutations in the 556 cancer-associated genes; this analysis identified seven additional events (those that were likely loss of heterozygosity related to copy number alterations in the tumor were not included), yielding a final list of 77 blood-specific mutations in 58 cases (Table 1 and Supplementary Table 5a). For five of those 58 cases that also had adjacent normal tissue analyzed, blood variants were absent in the adjacent normal tissue. Interestingly, among the 31 genes harboring these events, 19 have already been linked to hematological malignancies (Fig. 1a). More strikingly, 64 out of the total 77 events (83%) were in these 19 genes, examples as follows: *DNMT3A*<sup>16</sup> (18 cases), *TET2*<sup>17</sup> (9 cases), *JAK2*<sup>18</sup> (8 cases), *ASXL1*<sup>19,20</sup> (6 cases), *TP53*<sup>21</sup> (4 cases), *SF3B1*<sup>22</sup> (2 cases), *BCORL1*<sup>23</sup> (2 case), *ASXL2*<sup>24</sup> (1 case), and *SH2B3*<sup>25</sup> (1 case) (Table 1 and Fig. 1a). The overall frequency of blood-specific mutations increased with age (logistic regression analysis,  $P = 2.38e-08$ ), for example, 0.9% of the cases were in their 40s, 1.0% in their 50s, 1.8% in their 60s, 5.3% in their 70s, and 6.1% in their 80s (Fig. 1b). The blood-specific mutations were found in all 11 cancer types (Fig. 1b). Frequencies of individual *TET2*, *DNMT3A*, *ASXL1*, and *SF3B1* mutations also show association with age, (all FDR values  $< 0.034$ , logistic regression). Interestingly, *TET2*, *ASXL1*, and *SF3B1* mutations were predominantly found in the oldest age groups (70s and 80s) and *DNMT3A* mutations in 60s to 80s. In contrast, *JAK2* mutations, which did not achieve significance, trended in both younger (40s and 50s) and older age groups (70s) (Fig. 1c).

The average age of the 54 individuals with blood-specific mutations in the 19 leukemia/lymphoma-associated genes was  $70.0 \pm 9.9$  years, significantly higher than that of the larger TCGA cohort used in this study (difference of means test,  $P = 3.4e-10$ ) (Fig. 2a). Notably, the six individuals having two blood-specific mutations in the nine recurrently mutated genes are relatively older, with ages of 64 (*DNMT3A*: R882H; 36% VAF; *TET2*: H863fs; 12% VAF), 72 (*JAK2*: V617F; 73% VAF and *TET2*: T229fs; 19% VAF), 75 (*DNMT3A*: Y584fs; 38% VAF and *TET2*: Q764fs; 33% VAF), 76 (*JAK2*: V617F; 42% VAF and *ASXL1*: R548fs; 35% VAF), 83 years (*TET2*: F381fs; 50% VAF and *TET2*: Q888\*; 20% VAF), and unknown age (*BCORL1*: G883E; 17% VAF and *TP53*: Q136\*; 18% VAF), respectively (Table 1 and Supplementary Table 5a). We also compared the distribution of variant allele fractions for these 64 events versus inherited sites identified in the same sample set; we observed a clear shift towards lower VAFs in the blood-specific sites (Fig. 2b), suggesting the majority are present in only a fraction of the blood cells.

Although *GNAS* mutations have been found in leukemia<sup>26</sup>, activating gain-of-function mutations in *GNAS* are best known for their involvement in polyostotic fibrous dysplasia and McCune-Albright syndrome<sup>27</sup>. Interestingly, previous studies showed that activating mosaic *GNAS* mutations could affect various tissue types and the non-mosaic state for activating *GNAS* mutations may be lethal for the embryo<sup>28-30</sup>. This is consistent with our finding of mosaic *GNAS* R202H in transcript ENST00000354359 (also known as R201H in transcript ENSP00000360126) in the three blood samples of TCGA cases (11.5%, 14.4%, 21.4% VAFs, respectively). It is also worth noting that two blood-specific truncation mutations were detected in *PPMID*, a gene recently found to be associated with breast and ovarian predisposition with mosaic signatures<sup>31</sup>, but not with hematological malignancies.

Due to the fact that the blood-specific variants were present in only a fraction of the blood cells, we postulated that certain low level variants associated with clonal expansion were not captured by the variant detection tools. Therefore, we performed read count-based analysis for a set of known hotspot variants, including R882 in *DNMT3A*, R132 in *IDH1*, R172/R140 in *IDH2*, V617 in *JAK2*, K700 in *SF3B1*, and S34 in *U2AF1* for the entire TCGA cohort. We compared the distributions of VAFs between tumor and blood normal samples and found that the strongest differences were at R882 in *DNMT3A* and V617 in *JAK2* (Fig. 3 and Table 1). We devised a statistical procedure (see Methods) to identify additional “low VAF” sites significantly above the background error rate. Our analysis identified 14 blood samples having low-level hot spot variants (12 of them are not part of the 58 cases identified), including eight in *DNMT3A*, four in *JAK2*, one in both *SF3B1* and *IDH2*, but none in *IDH1* or *U2AF1*. The average age for these 14 cases is  $64.0 \pm 14.9$  years, again higher than the entire TCGA cohort studied (Supplementary Table 5b). We performed deep sequencing for five selected low-level hot spot variants (two R882C and one R882H sample sites in *DNMT3A*, one V617F in *JAK2*, and one K700E in *SF3B1*) to evaluate our detection approach. We achieved more than  $450,000\times$  average coverage for each site and all of the five variants were validated as bona fide low-level blood specific mutations (Supplementary Table 5c and see Methods). By including low VAF events, the overall frequencies of blood-specific mutations in different age groups are: 1.2%, 1.3%, 2.2%, 6.1%, and 6.8% in their 40s, 50s, 60s, 70s, and 80s, respectively (Supplementary Fig. 1). Collectively, we estimate that approximately 2% and 3.5% of individuals over the ages of 40 and 60, respectively, and without overt hematological malignancies carry blood-specific mutations that are associated with hematological malignancies.

### Known Hotspot Variants in NHLBI exome sequencing project control cohort

To confirm that these mutations are also present in an independent set of normal blood samples, we examined the NHLBI exome sequencing project (ESP) control cohort. We searched for RTVs and KHV in 6,503 ESP samples, focusing on four AML-associated genes discovered in the TCGA set (*DNMT3A*, *TET2*, *JAK2*, and *ASXL1*). When we applied a 0.1% MAF threshold in ESP (to focus on rare variants and to prevent potential false positives), we identified 8 RTVs and 13 KHV in *DNMT3A*, 13 RTVs in *TET2*, and 7 RTVs in *ASXL1*, and 3 KHV in *JAK2* (Supplementary Table 6). For a subset of ESP samples ( $n = 557$  WHISP), we re-aligned reads and performed variant analysis using the same process applied for the TCGA cohort. This allowed us to confirm RTVs and KHV detected by the pipeline, and also to detect low VAF KHV potentially missed by variant detection tools. Our pipeline detected one RTV each in *DNMT3A*, *TET2*, and *ASXL1*, four *DNMT3A* R882H mutations, and two *JAK2* V617F mutations. Careful analysis of recurrent hotspot sites (described in the previous section) using the 557 WHISP samples identified an additional three *DNMT3A* R882, one *JAK2* V617, one *GNAS* R202 and two *SF3B1* K700 variants with VAFs ranging from 2% to 10% (Fig. 3 and Supplementary Table 7). Based on these analyses, over 2% of the 557 WHISP cases contain mutations in these five selected genes. However, sequencing of matched non-blood samples from these cases would be required to prove that they are truly blood-specific mutations.

## Mutations in TCGA blood samples and patients with hematological malignancies

We next compared the mutation frequencies in 25 genes frequently mutated in at least one of the five following cohorts: TCGA 58 blood samples, 151 myeloproliferative neoplasm (MPN) cases reported by Nangalia *et al.*<sup>32</sup>, 150 myelodysplastic syndrome (MDS) cases reported by Walter *et al.*<sup>33</sup>, 160 chronic lymphocytic leukemia (CLL) by Landau *et al.*<sup>34</sup>, and 200 TCGA AML cases<sup>21</sup>. *DNMT3A*, *TET2*, *ASXL1*, *TP53*, and *SF3B1* were found to be consistently mutated in at least four groups, while *JAK2* was more specifically mutated in 58 TCGA blood samples and MPN patients (Fig. 4a). No mutations were found in TCGA blood samples in genes such as *IDH1*, *NRAS*, *RUNX1*, and *PHF6*, which are significantly mutated in AML and frequently mutated in MPN and/or MDS. Several genes showed cohort specificity: *GNAS* was mutated only in TCGA blood samples and *CEBPA*, *WT1*, *PTPN11*, *KIT*, *SMC1A*, and *SMC3* were preferably mutated in the AML cohort. We reasoned that common mutations among the cohorts (e.g., in *DNMT3A*, *TET2*, *ASXL1*, *SF3B1*, *JAK2*, and *TP53*) may be relevant for initiating HSPC clonal expansion, and are also likely to be early, initiation events for hematological malignancies, such as MPN, MDS, CLL, and AML. On the other hand, genes specific for MPN, MDS, and/or AML (e.g., *NRAS*, *RUNX1*, *NPM1*, and *FLT3*) are more likely to be subsequent, cooperating mutations that are involved in the progression of these diseases. These observations also show both distinct and common connections among these five cohort groups, suggesting that the TCGA cohort consists of a combination of precursor mutations that may sometimes evolve to MPN, MDS, CLL, and/or AML, although subsequent, collaborating mutations are clearly required (Supplementary Fig. 2). Finally, we compared the average number of mutations among these 4 cohorts (not including MDS) in 556 selected cancer-associated genes; this showed TCGA blood samples having the fewest mutations, MPN and CLL cohorts having intermediate numbers of mutations, and AML patients harboring the highest number of mutations (Fig. 4b).

## DISCUSSION

We identified age-related hematopoietic clonal expansion and the concurrent presence of leukemia/lymphoma-associated mutations in about 2% of individuals studied, who did not have reported hematological malignancies; this frequency reaches 5–6% for individuals who are at least 70 years of age. Our investigations of 2,728 TCGA blood samples identified 64 mutations in 19 genes known for their roles in hematological malignancies with VAFs above 10%, and an additional 14 mutations with lower VAFs (2 to 10%) when site-specific analysis was conducted. While many of these genes (e.g., *DNMT3A*, *JAK2*, *ASXL1*, and *TET2*) are established drivers for hematological malignancies, others (e.g., *PPM1D*) have not yet been implicated. The wide range of VAFs is indicative of the different stages of clonal expansion among individuals. Our finding also supports the hypothesis that mutations in genes such as *DNMT3A*, *JAK2*, *ASXL1*, *TET2*, *GNAS*, and others are likely to be initiating events for MPN, MDS, CLL, and/or AML, while epigenetic changes have also been previously implicated<sup>35,36</sup>. Importantly, the lack of detectable mutations in *IDH1*, *RUNX1*, *NRAS*, *NPM1*, and *FLT3* in both TCGA blood samples and WHISP cases supports the idea that these mutations are usually cooperating mutations that are important for disease progression.

These data suggest that extra care is required when using blood as a surrogate reference for the germline genome, especially in elderly individuals. First, there is an obvious risk of blood-specific variants in individuals without overt hematological malignancy being mistaken as germline variants. Secondly, germline alleles in cancer samples could be mistaken as tumor-specific variants when comparing to blood samples. Lastly, connections were made between mosaic *PPM1D* mutations in lymphocytes and the predisposition to breast and ovarian cancers<sup>31</sup>. While the influence of the immune system on tumorigenesis is well known, it is also possible that the blood-specific mutations are independent, unrelated events that are simply associated with the clonal expansion of HSPCs.

Our unbiased mutational analysis using sequencing data of relatively high depth (average of 107.5x coverage, Supplementary Table 8) allowed us to detect hot-spot mutations down to 2–3% VAFs; we discovered that 5–6% cases with advanced age (over 70 years) carry blood-specific mutations known to be involved in hematological malignancies. Some of these individuals may be undergoing hematopoietic clonal expansion (Fig. 5), but most probably do not progress to overt disease, since the incidence of myeloid malignancies in the elderly is less than 0.1%<sup>37</sup>. Participants providing specimens to TCGA are de-identified/coded, so it is not feasible to determine whether a participant with a leukemia-associated mutation actually progressed to a malignant hematologic disease (Supplementary Fig. 2). Using SNP array data from GENEVA study cohorts (melanoma, lung health, prostate cancer etc.), Laurie *et al.* showed roughly 2–3% of elderly individuals (over 70 years) have chromosomal anomalies in blood samples<sup>7</sup>. We therefore suggest that our estimate of frequency may be conservative, since other types of alterations (such as gene fusions and copy number alterations) were not included in our study. Regardless, these data clearly show that the elderly often acquire clonal “skewing” in their hematopoietic compartments and that this may represent a contributing factor to the development of hematologic malignancies.

## METHODS SUMMARY

We analyzed the peripheral blood sequence data from 2,728 individuals having had first-time primary cancer and no radiation or chemotherapy treatment. Exome data were aligned to human reference build 37 using BWA<sup>38</sup> and variants were identified using VarScan<sup>39</sup>, GATK<sup>40</sup>, and Pindel<sup>41</sup>, with stringent downstream filtering to standardize specificity. Variant annotation was based on Ensembl release 70\_37\_v5. The list of 556 cancer-associated genes was compiled from publicly available screening panels, published studies, and preliminary analysis of publicly available data sources<sup>11–15,42</sup>. Read count analysis was performed with our bam-readcount tool, available at <https://github.com/genome/bam-readcount>. Low level blood-specific events were discovered using a two-stage process including pre-filtering candidate non-mosaic samples using Bayes’ Rule and then the detection of high probability mosaic sites using Fisher’s exact test.

## METHODS

### Variant calling and annotation strategies

Exome sequencing data were aligned to NCBI Build 37 of the human reference using BWA v0.5.9 and de-duplicated using Picard 1.29. Single nucleotide variants were identified by

Varscan (version 2.2.6: `-min-var-freq 0.1 -p-value 0.1 -min-coverage 8 -map-quality 10`), and GATK (revision 5336: `-T UnifiedGenotyper -R GRCh37-lite -et NO_ET -I INFO -U ALL -validation_strictness SILENT`). Indels were identified using Varscan (version 2.2.9: `-min-coverage 3 -min-var-freq 0.2 -p-value 0.1 -strand-filter 1 -map-quality 10`), GATK (revision 5336: `-T IndelGenotyperV2 -R GRCh37-lite -window_size 300 -et NO_ET -U ALL`), as well as Pindel (version 0.2.4×, May 8, 2013: `-window-size 1`). For Pindel analysis, we preset the insertion size to 500 if this information is not provided in the BAM header. SNVs were based on the union of GATK and VarScan. They were subsequently processed through our in-house false-positive filter (`-min-homopolymer 10`). We required that indels were called by at least 2 out of 3 callers (GATK, Varscan, Pindel). In addition we also included Pindel unique calls (at least 30× coverage and 20% VAF). All combined indels were then processed through our false-positive-filter (`-min-homopolymer 10 -min-var-freq 0.2 -min-var-count 6`). We then applied additional annotation and minor allele frequency filters as previously reported<sup>10</sup>.

Variant transcript annotation is based on all human transcripts obtained from Ensembl Release 70\_37\_v5. The reference alleles and positions were derived from the sequence and coordinates of GRCh37. All transcripts were annotated and a single representative was selected for each somatic mutation based on the significance of the predicted functional effect of each mutation, ordered from most significant to least significant as follows: nonsense, frameshift, splice site, in frame, missense, no stop (nonstop/readthrough), synonymous, and RNA. Splice site mutations were restricted to substitutions, deletions, or insertions overlapping the 2bp intronic sequence defined as the canonical splice donor or splice acceptor. RNA mutations were restricted solely to transcripts without an annotated open reading frame. Mutations affecting 3'UTR, 5'UTR, intronic sequence, and intergenic sequence were discarded for the purposes of downstream analysis.

### Recurrent somatic mutations in 12 cancer types

We collected extensively filtered somatic variants in 3,355 TCGA samples from 12 cancer types (Supplementary Table 3a), and selected recurrent mutations appearing more than once at a given genomic position (Supplementary Table 3b).

### Compiling cancer-associated gene list

A total of 556 candidate cancer-associated genes were compiled using nine sources, including recently published large-scale cancer studies, publicly available screening panels, and unpublished preliminary analysis of publicly available data sources. The 204 genes shared across at least two of the nine sources were retained and a literature search was conducted to identify evidence supporting inclusion of any remaining unique genes. A subset of 518 genes originated from recent publications, including 294 genes from Frampton *et al.*<sup>15</sup>, 125 genes from Kandoth *et al.*<sup>11</sup>, 212 genes from Lawrence *et al.*<sup>13</sup>, 194 genes from Pritchard *et al.*<sup>14</sup>, and 124 genes from Vogelstein *et al.*<sup>12</sup> Thirty-nine additional genes were included based on the analysis of driver mutations in 20 TCGA cancer types (unpublished), recommendations in accordance with the standards and guidelines of the American College of Genetics and Genomics<sup>42</sup>, and 18 novel cancer driver genes identified in recently published large-scale studies (Supplementary Table 4).



## Readcount analysis and statistical approaches for identifying significant low level variants

Read counts for variants were determined using our internally developed tool bam-readcount (<https://github.com/genome/bam-readcount>). For sites to be included in the downstream statistical test, we require greater than 30× coverage for both blood normal and matched tumor samples. We assessed the false-discovery rate (FDR) using a two-stage process. The first step is a purpose-specific pre-filter to eliminate candidates that can be confidently identified as having originated from non-mosaic samples, as these identically fail the inclusion criterion for significance testing. It targets deeply covered sites whose apparent variant reads actually represent base-calling errors. Take the sample space for blood classification as consisting of two mutually exclusive, collectively exhaustive (MECE) statuses, “normal”,  $S=N$ , and “mosaic”,  $S=M$ , and let the candidate blood site data,  $D$ , consist of  $T$  spanning reads, of which  $V$  and  $T-V$  report variant and reference counts, respectively. If we presume that the rate of mosaicism (the fraction of altered cells, assumed as roughly 2%),  $\rho$ , is much larger than the Phred-determined likelihood of base-calling error for any single read,  $\varepsilon$ , i.e.  $\rho/\varepsilon \gg 1$ , then the conditional probabilities are binomial,  $P(D|S) = C_{T,V}^T p^V (1-p)^{T-V}$ , where  $C_{T,V}$  is the number of combinations of  $T$  objects chosen  $V$  at a time and  $p=\varepsilon$  for  $S=N$  and  $p=\rho$  for  $S=M$ . Given a prior estimate of  $P(N)=0.999$ , we can formulate  $P(N|D)$  directly from Bayes’ Rule, from which additional algebraic manipulation and suitable asymptotic approximation show

$$P(N|D) = \frac{1}{1 + P(M) \cdot (\rho/\varepsilon)^V \cdot \exp[-\rho(T-V)] / P(N)}$$

Candidates are culled as non-mosaic if this probability exceeds 95%, though in practice most cases actually removed have  $P(N|D) > 99.5\%$ . The remaining set of events is subsequently passed to the second step, which is a traditional Fisher exact (table) test for association between sample type and variant allele fraction followed by standard Benjamini-Hochberg FDR assessment. We report events having 20% FDR with at least 3 supporting reads and greater than 2% variant allele fraction.

### Analysis of NHLBI ESP data

NHLBI variants calls for 6,503 samples were downloaded from the NHLBI Exome Variant Server <http://evs.gs.washington.edu/EVS/>. All variants were processed using the same tools as for the TCGA cohort. For comparative analysis, all ESP variants are filtered for < 0.1% total MAF to minimize false positives. The Women’s Health Initiative Exome Sequencing Project (WHISP) is part of the National Heart, Lung, and Blood Institute’s (NHLBI), Grand Opportunity Exome Sequencing Project (<https://www.fhrc.org/en/labs/phs/projects/cancer-prevention/projects/whisp.html>). WHISP data for 614 samples were downloaded from dbGaP, verified for file integrity, and then imported as BAM files into our data warehouse. Alignment to the reference genome GRCh37-lite was carried out using BWA v0.5.9 with parameters  $-t\ 4\ -q\ 5$  and marking of duplicates by Picard v1.46. Variant calling was performed as described in the "Variant calling and annotation strategies". For quality control purposes, we included WHISP samples with read mapping rates greater than 80%, duplication rates less than 40%, and at least 10,000 SNVs detected in the target region. The

557 Caucasian WHISP samples selected for this study, on average, had mapping rates of ~95%, duplication rates of ~9%, and ~18,000 SNVs called in the target region.

### Deep sequencing validation

Five candidate low-level variants (two R882C and one R882H sample sites in *DNMT3A*, one V617F in *JAK2*, and one K700E in *SF3B1*), 5 positive controls, and 4 negative controls were selected for validation using deep sequencing (Supplementary Table 5c). Primer pairs tailed with sample-specific indexes were designed for individual target sites and further used for PCR amplifications. Indexed libraries were made for tumor and blood pools respectively. We then generated sequencing data using 1 lane of MiSeq run with read length of 2×250. Custom references were created by including specific primer and index sequences. MiSeq reads were aligned to the custom references using BWA (bwa aln -t 8; bwa sampe). Allelic counts for the variants were obtained using in house tool bam-readcount (bam-readcount -b 0 -q 30).

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

### Acknowledgments

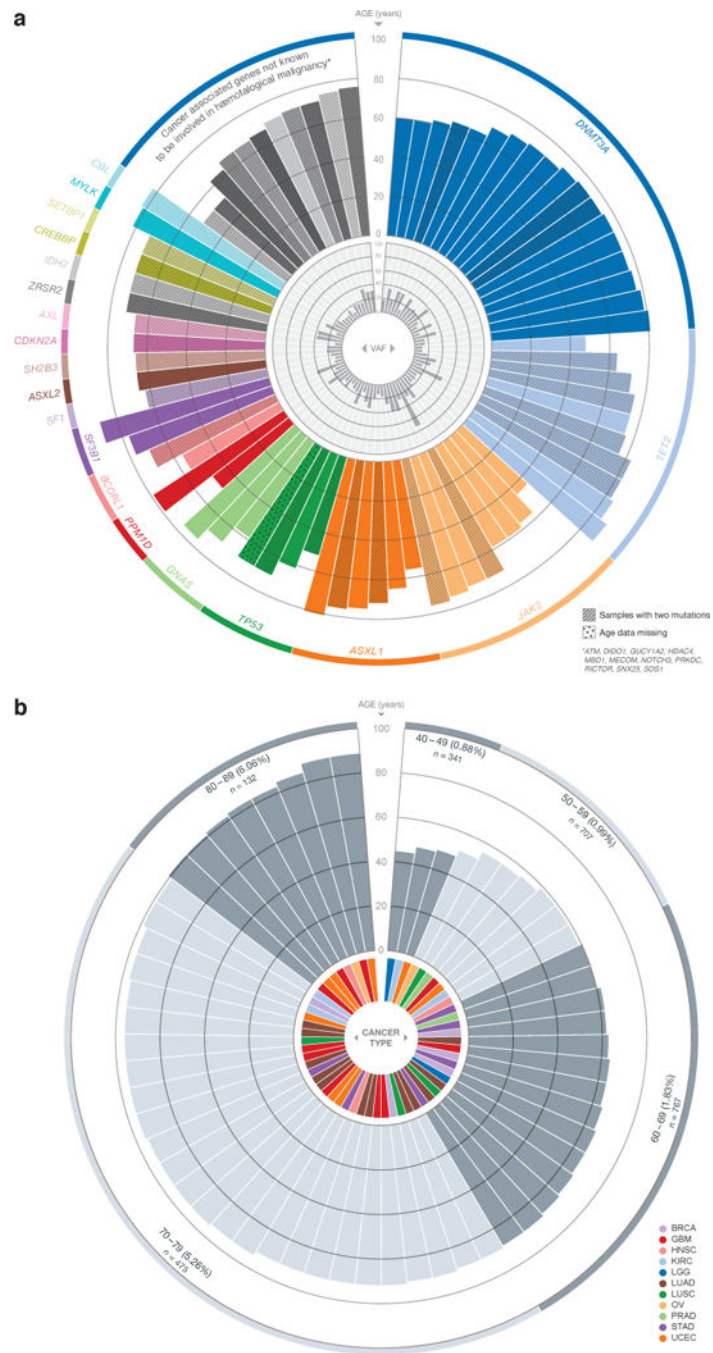
This work was supported by the US National Cancer Institute grants R01CA180006 to L.D. and P01CA101937 to T.J.L. and US National Human Genome Research Institute grants U54HG003079 to R.K.W. and U01HG006517 to L.D. M.J.W. is supported by Leukemia and Lymphoma Society Scholar Award (1230–14) and J.S.W. is supported by R00HL103975. We thank M. Wyczalkowski for suggestions on figures and C. Kandath on somatic mutation analysis. We also acknowledge The Cancer Genome Atlas (cancergenome.nih.gov) as the source of primary data.

### References

1. Welch JS, et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*. 2012; 150:264–278. [PubMed: 22817890]
2. Limpens J, et al. Lymphoma-associated translocation t(14;18) in blood B cells of normal individuals. *Blood*. 1995; 85:2528–2536. [PubMed: 7727781]
3. Liu Y, Hernandez AM, Shibata D, Cortopassi GA. BCL2 translocation frequency rises with age in humans. *Proc Natl Acad Sci U S A*. 1994; 91:8910–8914. [PubMed: 8090743]
4. Biernaux C, Loos M, Sels A, Huez G, Stryckmans P. Detection of major bcr-abl gene expression at a very low level in blood cells of some healthy individuals. *Blood*. 1995; 86:3118–3122. [PubMed: 7579406]
5. Forsberg LA, et al. Age-related somatic structural changes in the nuclear genome of human blood cells. *Am J Hum Genet*. 2012; 90:217–228. [PubMed: 22305530]
6. Jacobs KB, et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat Genet*. 2012; 44:651–658. [PubMed: 22561519]
7. Laurie CC, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet*. 2012; 44:642–650. [PubMed: 22561516]
8. Busque L, et al. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat Genet*. 2012; 44:1179–1181. [PubMed: 23001125]
9. Shlush LI, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*. 2014; 506:328–333. [PubMed: 24522528]
10. Kanchi KL, et al. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature communications*. 2014; 5

11. Kandoth C, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013; 502:333–339. [PubMed: 24132290]
12. Vogelstein B, et al. Cancer genome landscapes. *Science*. 2013; 339:1546–1558. [PubMed: 23539594]
13. Lawrence MS, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014; 505:495–501. [PubMed: 24390350]
14. Pritchard CC, et al. Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens. *The Journal of molecular diagnostics : JMD*. 2014; 16:56–67. [PubMed: 24189654]
15. Frampton GM, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol*. 2013; 31:1023–1031. [PubMed: 24142049]
16. Ley TJ, et al. DNMT3A mutations in acute myeloid leukemia. *N Engl J Med*. 2010; 363:2424–2433. [PubMed: 21067377]
17. Delhommeau F, et al. Mutation in TET2 in myeloid cancers. *N Engl J Med*. 2009; 360:2289–2301. [PubMed: 19474426]
18. Kralovics R, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N Engl J Med*. 2005; 352:1779–1790. [PubMed: 15858187]
19. Gelsi-Boyer V, et al. Mutations of polycomb-associated gene ASXL1 in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br J Haematol*. 2009; 145:788–800. [PubMed: 19388938]
20. Abdel-Wahab O, et al. Deletion of Asxl1 results in myelodysplasia and severe developmental defects in vivo. *J Exp Med*. 2013; 210:2641–2659. [PubMed: 24218140]
21. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N Engl J Med*. 2013
22. Yoshida K, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*. 2011; 478:64–69. [PubMed: 21909114]
23. Li M, et al. Somatic mutations in the transcriptional corepressor gene BCORL1 in adult acute myelogenous leukemia. *Blood*. 2011; 118:5914–5917. [PubMed: 21989985]
24. Zhang J, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*. 2012; 481:157–163. [PubMed: 22237106]
25. Perez-Garcia A, et al. Genetic loss of SH2B3 in acute lymphoblastic leukemia. *Blood*. 2013; 122:2425–2432. [PubMed: 23908464]
26. Bejar R, et al. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med*. 2011; 364:2496–2506. [PubMed: 21714648]
27. Weinstein LS. G(s)alpha mutations in fibrous dysplasia and McCune-Albright syndrome. *Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research*. 2006; 21(Suppl 2):120–124.
28. Aldred MA, Trembath RC. Activating and inactivating mutations in the human GNAS1 gene. *Hum Mutat*. 2000; 16:183–189. [PubMed: 10980525]
29. Lumbroso S, Paris F, Sultan C. Activating Gsalpha mutations: analysis of 113 patients with signs of McCune-Albright syndrome—a European Collaborative Study. *J Clin Endocrinol Metab*. 2004; 89:2107–2113. [PubMed: 15126527]
30. Ringel MD, Schwindinger WF, Levine MA. Clinical implications of genetic defects in G proteins. The molecular basis of McCune-Albright syndrome and Albright hereditary osteodystrophy. *Medicine*. 1996; 75:171–184. [PubMed: 8699958]
31. Ruark E, et al. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature*. 2013; 493:406–410. [PubMed: 23242139]
32. Nangalia J, et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N Engl J Med*. 2013; 369:2391–2405. [PubMed: 24325359]
33. Walter MJ, et al. Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia*. 2013; 27:1275–1282. [PubMed: 23443460]

34. Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*. 2013; 152:714–726. [PubMed: 23415222]
35. Akalin A, et al. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet*. 2012; 8:e1002781. [PubMed: 22737091]
36. Lu C, et al. Induction of sarcomas by mutant IDH2. *Genes Dev*. 2013; 27:1986–1998. [PubMed: 24065766]
37. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. *CA Cancer J Clin*. 2013; 63:11–30. [PubMed: 23335087]
38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–1760. [PubMed: 19451168]
39. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012; 22:568–576. [PubMed: 22300766]
40. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20:1297–1303. [PubMed: 20644199]
41. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z, Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009; 25:2865–2871. [PubMed: 19561018]
42. Rehm HL, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics*. 2013; 15:733–747. [PubMed: 23887774]



**Figure 1. Blood-specific mutations identified in 58 out of 2,728 TCGA cases from 11 cancer types**  
**(a)** Rose chart illustrating the distribution of blood-specific nonsynonymous mutations in 31 genes. The variant allele fractions (VAF) of the 77 mutations are indicated in the center. **(b)** Rose chart illustrating the age distribution of samples with blood-specific mutations. Higher frequencies of blood-specific mutations are found in older age groups (60s, 70s, and 80s) versus younger ones (40s and 50s). The cancer type distribution is shown in the center. **(c)** Distribution of blood-specific mutations in *DNMT3A*, *TET2*, *JAK2*, *ASXL1*, *SF3B1*, and

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

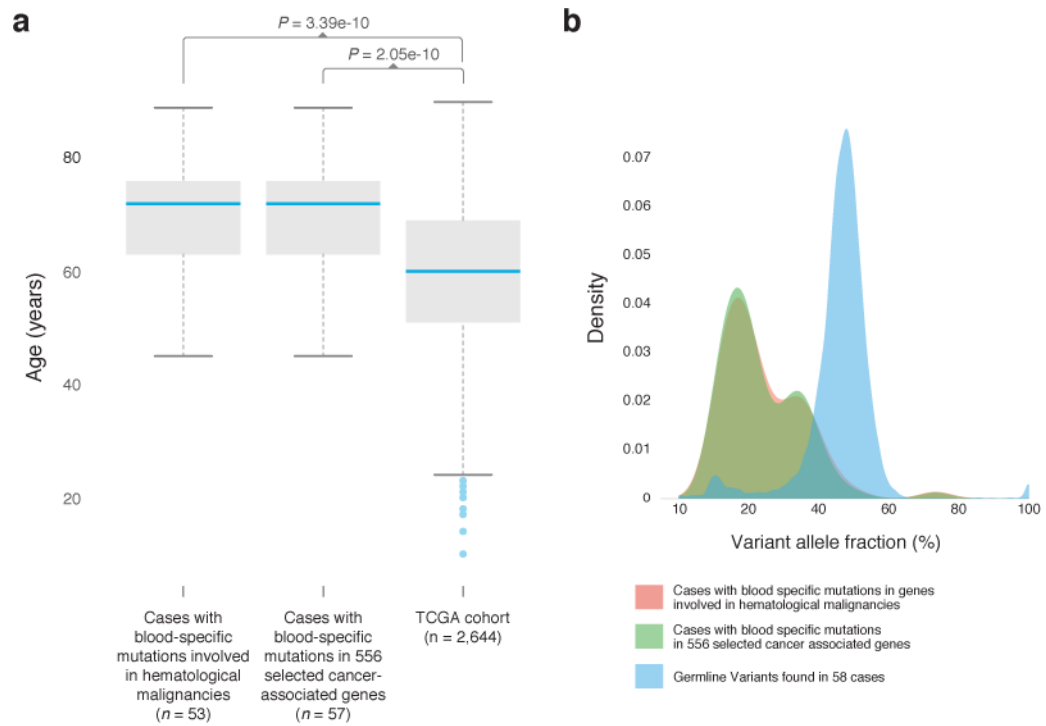
*GNAS* in different age groups. Total includes all blood-specific mutations in 556 cancer associated genes identified in each age group.

Author Manuscript

Author Manuscript

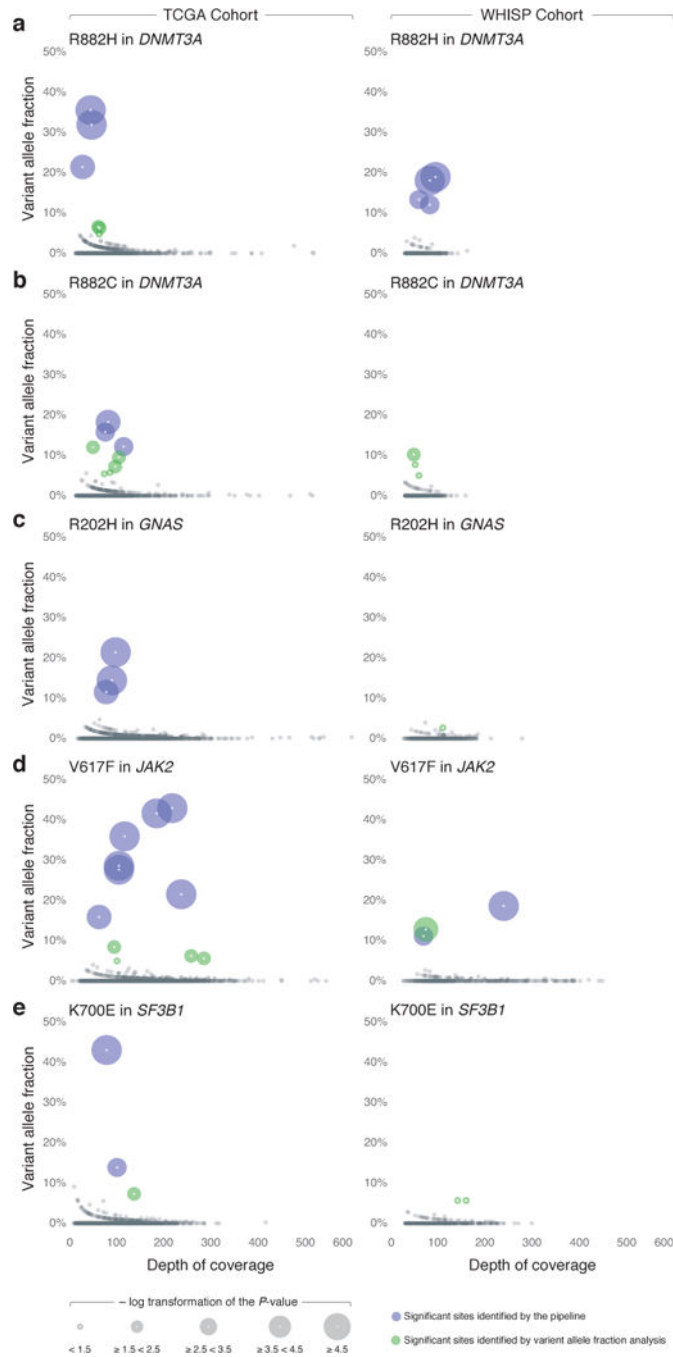
Author Manuscript

Author Manuscript



**Figure 2. Blood-specific mutations and their association with age**

(a) Box plot showing positive correlation between blood-specific mutations in leukemia/lymphoma genes and age. Age information is not available for one of the 58 cases with blood specific mutations. (b) The wide spectrum and lower average variant allele fractions in blood-specific mutations, compared to the ~50% VAF for germline variants identified in the same samples.



**Figure 3. Low VAF blood-specific, hotspot mutations identified in the TCGA and WHISP cohorts using a readcount based approach**

Blood-specific mutations identified by the variant detection pipeline are in blue. An additional 14 blood-specific events (13 shown in green) with VAFs between 2% and 10% were identified in the TCGA samples and their positive associations with older ages were confirmed. 13 hotspot variants were identified in WHISP samples ( $n = 557$ ) and seven (in green) have variant allele fractions ranging from 2% to ~10%. One *JAK2* V617F identified



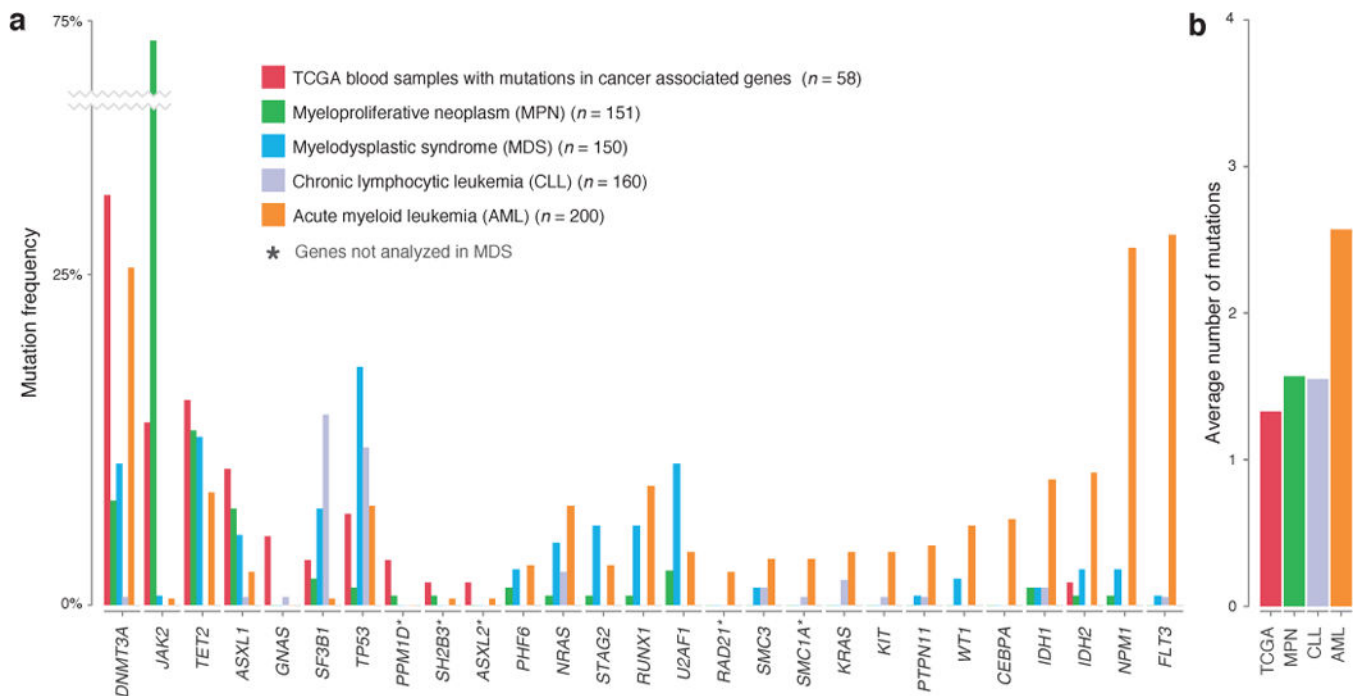
in a TCGA sample was not shown due to a VAF higher than 50%. **(a)** *DNMT3A* R882C, **(b)** *DNMT3A* R882H, **(c)** *GNAS* R202H, **(d)** *JAK2* V617F, and **(e)** *SF3B1* K700E.

Author Manuscript

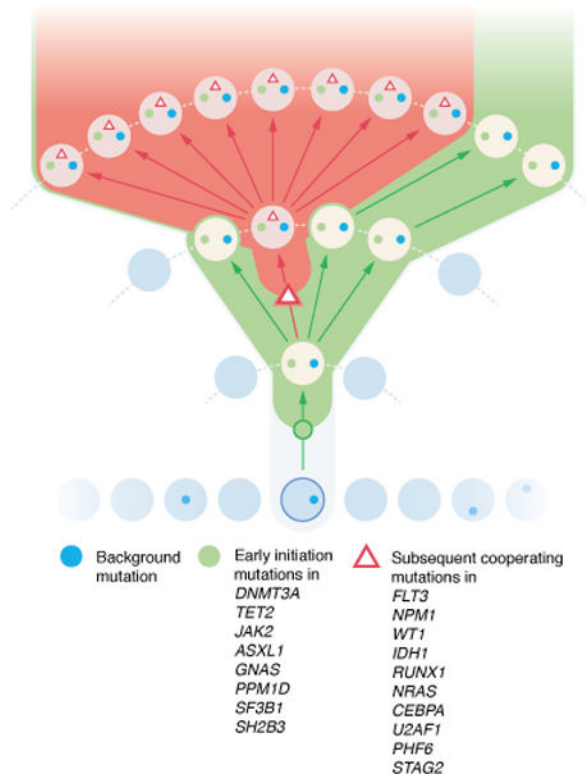
Author Manuscript

Author Manuscript

Author Manuscript



**Figure 4. Comparison of mutation frequencies in blood samples from 58 TCGA cases with mutations in cancer-associated genes in 151 MPN, 150 MDS, 160 CLL, and 200 AML cases (a) Mutation frequencies of major genes involved in hematological malignancies. (b) The average number of non-synonymous mutations found in TCGA blood normal cases, MPN, MDS, and AML patients across 556 cancer associated genes.**



**Figure 5. Clonal expansion model**

The distinct roles of a set of genes including *DNMT3A*, *ASXL1*, *TET2*, *GNAS*, *JAK2*, *PPM1D*, *IDH1*, *NRAS*, *NPM1*, and *FLT3* in the initiation of hematopoietic clonal expansion.

**Table 1**  
**Blood-specific mutations in 9 recurrently mutated genes identified in TCGA cases**

Asterisks indicate nonsense mutations. VAF is defined as the proportion of reads supporting the variant allele. 11 cancer types were investigated in this study: BRCA (breast adenocarcinoma), GBM (glioblastoma multiforme), HNSC (head and neck squamous cell carcinoma), KIRC (kidney renal clear cell carcinoma), LGG (brain low grade glioma), LUAD (lung adenocarcinoma), LUSC (lung squamous cell carcinoma), OV (ovarian carcinoma), PRAD (prostate adenocarcinoma), STAD (stomach adenocarcinoma), and UCEC (uterine corpus endometrial carcinoma).

Gene	Mutation	Case			Gene	Mutation	Case			
		Type	Age	VAF			Type	Age	VAF	
<i>DNNMT3A</i>	p.R882C	GBM	81	15.79%	<i>JAK2</i>	p.V617F	GBM	57	21.52%	
		STAD	60	18.29%			GBM	72	73.39%	
	STAD	69	12.17%	KIRC		59	28.57%			
	BRCA	62	21.43%	LGG		45	15.87%			
	p.R882H	GBM	64	35.56%		LUAD	72	27.62%		
		LUSC	76	31.91%		LUAD	76	41.62%		
	e13+1	KIRC	79	15.94%		UCEC	59	35.90%		
		LUAD	76	11.11%		UCEC	74	42.92%		
	p.E469*	GBM	72	20.60%		<i>ASXL1</i>	p.Q733*	LUAD	75	20%
	p.F851fs	BRCA	64	34.88%				LUAD	72	14.29%
p.K577fs	HNSC	72	24.14%	p.Q733fs	UCEC		81	27.27%		
p.N516fs	LUSC	71	33.33%	p.R548fs	LUAD		76	35.03%		
p.S770*	STAD	75	16.03%	p.Y591*	STAD		65	17.88%		
p.W314*	UCEC	74	22.06%	p.Y591fs	LUSC		56	29.70%		
p.Y584fs	GBM	75	38%	p.C275Y	OV		52	14.29%		
e12-1	PRAD	60	35.79%	<i>TP53</i>	p.Q136*		LUAD	null	18%	
	e21-2	GBM	76		11.81%		p.Q144*	STAD	62	15.96%
e22-1	UCEC	77	33.85%		p.R273L		LUAD	70	34.62%	
	GBM	83	50%		p.R202H	GBM	76	14.44%		
p.F381fs	GBM	64	11.67%			<i>GNAS</i>	HNSC	59	11.54%	
p.H863fs	GBM	64	11.67%		<i>PPM1D</i>	LUAD	69	21.43%		
p.K889*	OV	85	15.09%			p.Q520*	BRCA	79	35.42%	
p.Q531*	KIRC	48	11.90%		p.S468*	UCEC	49	21.23%		
p.Q644*	UCEC	89	16.78%							

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Gene	Mutation	Case			Gene	Mutation	Case		
		Type	Age	VAF			Type	Age	VAF
	p.Q764fs	GBM	75	33.01%	<b>BCORLI</b>	p.G883E	LUAD	null	16.67%
	p.Q831fs	LUAD	75	26.42%		p.S264*	PRAD	56	22.45%
	p.Q888*	GBM	83	20.39%	<b>SF3BI</b>	p.K700E	GBM	89	13.86%
	p.R550*	LUAD	76	16.25%			KIRC	77	43.04%
	p.T229fs	GBM	72	19.05%					