Supplementary Appendix

Supplement to: Genovese G, Kähler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med 2014;371:2477-87. DOI: 10.1056/NEJMoa1409405

Table of Contents

Supplementary Methods..................................2

Research participants and whole-exome sequencing of blood cell-derived DNA........2

Definition of putative, inclusive, and candidate driver somatic mutations.............3

Molecular validation of putative and candidate driver somatic mutations.............4

DNMT3A mutations....................................5

Somatic loss-of-Y chromosome...................5

Relationship between schizophrenia, smoking, and clonal hematopoiesis.............6

High coverage whole-genome sequencing of blood cell-derived DNA...............................6

Whole-exome and whole-genome sequencing of bone marrow biopsies...........7

Subject #1...8

Subject #2...8

Subject #3...8

Statistics and figures....................................9

Figures..10

Figure S1...10

Figure S2...11

Figure S3...12

Figure S4...13

Figure S5...14

Figure S6...15

Figure S7...16

Figure S8...17

Figure S9...18

Figure S10...19

Figure S11...20

Figure S12...21

Figure S13...22

Figure S14...23

Figure S15...24

Figure S16...25

Tables...26

Table S1...26

Table S2...27

Table S3...28

Table S4...36

Table S5...37

Table S6...38

Table S7...39

Table S8...40

Table S9...42

Table S10...43

Table S11...44

References..45

Supplementary Methods

Research participants and whole-exome sequencing of blood cell-derived DNA

A total of 12,380 Swedish research participants with psychiatric diagnoses were ascertained from the Swedish National Hospital Discharge Register, which captures all inpatient hospitalizations. Controls were randomly selected from population registers. We treated cases and controls as a single cohort for all analyses presented below, as none of the mutational variables analyzed below showed any relationship to psychiatric diagnosis after controlling for other factors such as age and smoking. Research participation and DNA sampling took place from 2005 to 2013.

Excluding bipolar subjects, medical histories (from 1965 to 2011) of 11,164 of the subjects enrolled in the study were extracted from the Swedish national in- and outpatient register (median follow-up was 32 months). Information about vital status (from 2006 to 2012) was extracted from the population register and the Cause of Death register (median follow-up was 42 months). To identify individuals with hematologic malignancies, we included diagnoses within ICD10 code groups C81–C96 (malignant neoplasms of lymphoid, hematopoietic and related tissue), D45 (polycythemia vera), D46 (myelodysplastic syndromes), D47 (other neoplasms of uncertain behavior of lymphoid, hematopoietic and related tissue), and D7581 (myelofibrosis) and the same diagnoses within the corresponding ICD9 and ICD8 groups.

The 12,380 samples collected were sequenced in twelve separate waves. The first wave employed an earlier version of the hybrid-capture procedure (Agilent SureSelect Human All Exon Kit), which targets ~28 million base pairs of the human genome, partitioned in ~160,000 intervals, whereas the samples from the other waves used a newer version (Agilent SureSelect Human All Exon v.2 Kit), which targets ~32 million base pairs of the human genome, partitioned in ~190,000 intervals. The first wave was sequenced using Illumina GAII instruments and the remaining waves were sequenced using Illumina HiSeq 2000 and HiSeq 2500 instruments, with pair ended sequencing reads of 76 base pairs across all waves. Sequencing was performed at the Broad Institute of MIT and Harvard across the period of time from 2010 to 2013.

Sequencing data were aligned against the GRCh37 human genome reference using BWA ALN version 0.5.9.[1] On average across samples each base pair of the target intervals was observed 95 times. Genotypes and allelic counts were computed across the genome using the Haplotype Caller from the Genome Analysis Toolkit version 3.1-1,[2] which generated genotypes for 1,812,331 variant sites across 12,380 subjects. Due to the specific default parameters used by the Haplotype Caller and aimed at genotyping inherited mutations, we recognized that several mutations present in sequencing reads in the 5-10% allele fraction range, and that could have been called, were not reported. To mitigate this issue, we used the Unified Genotyper from the Genome Analysis Toolkit to genotype 208 variants reported as seen seven or more times in hematopoietic or lymphoid cancers in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database[3] v69 (released June 2nd, 2014), with the exception of a few that we deemed inherited mutations or PCR sequencing artifacts[4] rather than somatic events (**Table S2**). We kept all mutations for which the alternate allele was observed on at least three sequencing reads in an individual's sequencing data. These thresholds yielded 26 additional mutations that were not called by the Haplotype Caller. We did not use these mutations for our unbiased analysis of enrichment of disruptive mutations.

Definition of putative, inclusive, and candidate driver somatic mutations

Due to the higher likelihood of misalignment and PCR artifacts, we excluded from analysis somatic mutations in the following regions:
1) Low complexity regions and sites harboring markers failing Hardy Weinberg equilibrium tests in the 1000 Genomes Project phase 1[5] (https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs37d5.bed.gz and https://github.com/lh3/varcmp/blob/master/scripts/1000g.hwe-bad.bed)
2) Sites with excess coverage within the 1000 Genomes Project phase 1[6]
3) Segmental duplications of the human genome[7,8] (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/genomicSuperDups.txt.gz)
4) Regions harboring common large insertions in 1000 Genomes Project Phase1 samples (data unpublished)
5) Regions excluded from the strict mask of the 1000 Genomes Project phase 1[9] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/20120824_strict_mask.bed)
These filters defined regions covering ~60% of the GRCh37 human genome reference and ~70% of the coding regions and they excluded 161,158 out of the 1,812,331 variants called in the cohort.

Due to enrichment bias in exome libraries, allelic fractions for inherited heterozygous mutations are not expected to be centered around 50% . The average expected allelic fraction for the alternate allele of a heterozygous single nucleotide polymorphisms (SNPs) is actually 47%±4% (**Fig. S2**). For indels, this value is even lower, likely due to a mix of enrichment bias, sequence misalignment, and improper reporting of allelic counts for some complicated indels from the Haplotype Caller from the Genome Analysis Toolkit prior to version 3.2 (**Fig. S3A,B**). Therefore we decided to apply different thresholds for SNPs and indels for the purpose of identifying putative somatic mutations.

We define as putative somatic mutations those alleles satisfying the following criteria:
1) SNPs
2) Observed once or twice (minor allele frequency less than 0.01%) in the cohort
3) Allelic fraction above 10%
4) Failed the hypothesis that the alternate allelic count was distributed as a binomial process with mean 45% with a designed false positive rate of 10^{-5}

We define as inclusive somatic mutations those alleles satisfying the following criteria:
1) SNPs or indels of length one or two base pairs
2) Observed at most six times (minor allele frequency less than 0.025%) in the cohort
3) Allelic fraction above 5%
4) Failed the hypothesis that the alternate allelic count was distributed as a binomial process with mean 47% for SNPs and 40% for indels with a designed false positive rate of 0.01

These definitions yielded 4,275 putative somatic mutations and 53,474 inclusive somatic mutations across 12,380 subjects. Upon further analysis, a large fraction of these mutations originated from the first two sequencing waves (**Fig. S4A,B**). This likely reflected older capture and sequencing technologies used during the first two waves. We also observed a single outlier subject from the sixth sequencing wave, with 193 putative somatic mutations and 1,207 inclusive somatic mutations. Putative somatic mutations from this outlier failed to validate in an independent experiment.

We excluded the 534 subjects from the first two waves and the outlier subject from any subsequent analyses in which putative or inclusive somatic mutations were used. This resulted in a refined set of 3,111 putative somatic mutations and 42,282 inclusive somatic mutations from 11,845 subjects. Mutational profiles for inherited mutations (**Fig. S5A**) resemble mutational profiles for inclusive and putative somatic mutation sets (**Fig. S5B,C**) suggesting that technical artifacts, rather than genuine somatic and inherited mutations, must constitute a small fraction of the two sets. By contrast, the mutational profiles for inclusive somatic mutations from the first two sequencing waves (**Fig. S5D,E**) were quite different, and so were the mutational profiles for inclusive somatic mutations in the outlier subject from the sixth sequencing wave (**Fig. S5F**), further suggesting that these were library preparation or sequencing artifacts rather than real biological events.

Finally, we define as candidate driver somatic mutations those alleles satisfying the following criteria:
1) Disruptive and missense mutations in gene *DNMT3A* localized in exons 7 to 23
2) Disruptive mutations in gene *ASXL1* with the exclusion of *ASXL1* p.G646fsX12 and p.G645fsX58
3) Disruptive mutations in gene *TET2*
4) Disruptive mutations in gene *PPM1D*
5) Missense mutation *JAK2* p.V617F
6) Mutations reported at least seven times in hematopoietic and lymphoid malignancies using the Catalogue of Somatic Mutations in Cancer[3] with the exclusions of inherited mutations and potential PCR artifacts (**Table S2**)
Notice that this definition does not take allelic fractions into account.

Due to low coverage in one small region of *ASXL1* (**Fig. S6B**) we were not able to discern mutation *ASXL1* p.G646fsX12, known to account for >50% of mutations in *ASXL1* in myeloid malignancies, from potential PCR artifacts.[4] Moreover the exome enrichment reagent we used does not capture some exons of *TET2* accounting for almost half of the coding region in which other studies have identified mutations[10] (**Fig. S6C**). Therefore mutations in *TET2* and *ASXL1* were likely under-ascertained in this study.

Molecular validation of putative and candidate driver somatic mutations

We performed a validation experiment for 65 mutations selected among putative somatic mutations and candidate driver somatic mutations from 12 subjects. A library preparation method utilizing a two-round tailed amplicon PCR strategy was used to create targeted sequencing libraries for sequencing at high coverage on an Illumina MiSeq instrument. Alignment of sequencing reads against the GRCh37 human genome reference was performed using BWA MEM version 0.7.7[11] and allelic fractions were computed using the Unified Genotyper from the Genome Analysis Toolkit version 3.2-2.[2] Each putative somatic mutation that we attempted to validate was confirmed as somatic (**Fig. S7**).

We further performed validation for 30 candidate driver somatic mutations from two well-known recurrently mutated sites, *DNMT3A* p.R882H and *JAK2* p.V617F. These were genotyped using TaqMan fluorescent assays in a droplet-based digital PCR system.[12] Relative concentrations of each allele were quantitated through multiplexed fluorophores counted across approximately 15,000 nanoliter-sized droplets. Each somatic mutation that we attempted to validate was confirmed as somatic, including five *JAK2* p.V617F mutations mutations showing at allelic fractions close to or above 50% (**Fig. S8**), as would be expected as a consequence of a loss-of-heterozygosity event.[13]

DNMT3A mutations

A total of 190 mutations across 185 subjects were identified in the *DNMT3A* gene (**Table S4**). Studies of mutations in hematologic malignancies have found *DNMT3A* mutations to be more common in cancers from females than in cancers from males.[14–16] We found that *DNMT3A* somatic mutations were also more common in females than in males (104/5780 vs. 81/6600; P=0.016 after adjusting for age using a linear regression model).

We observed 48 disruptive mutations, and 142 in-frame indels or missense mutations including 23 mutations affecting the R882 aminoacid of which 15 are R882H mutations known to dominantly inhibit wild-type *DNMT3A*.[17] We also observed an enrichment within the *DNMT3A* FF interface region bounded by amino acid F732 and amino acid F772,[18] similarly to what seen in *DNMT3A* mutations in acute myeloid leukemia (see http://cancergenome.broadinstitute.org/index.php?gene=DNMT3A).[19]

Of the 20 missense mutations within the FF interface region, 10 generated new cysteine residues (**Fig. S9A,B**). We posited that these new cysteine residues might inactivate *DNMT3A* protein function by inappropriately forming disulfide bonds if the protein were exposed to oxidizing environment during its biogenesis or function. We then used the DiANNA disulfide bond prediction tool[20] to predict disulfide bond formation for each of the mutant proteins containing a new cysteine residue. Out of 10 different cysteine forming mutations, 8 were predicted to form new disulfide bonds to other native cysteine residues located in the ADD, cysteine-rich, catalytic domain of *DNMT3A*[21] which spans amino acids 472-610 with high prediction scores (0.85±0.24, mean±S.D.) (**Table S4**). We then used a three-dimensional structure prediction tool[22] and were able to predict 51% of *DNMT3A* sequence (from R476 to F909), including the catalytic domain as well as the FF and RD domains, which are required in oligomerization of *DNMT3A*. Based on the three-dimensional structure of *DNMT3A*, most of the predicted *de novo* disulfide bonds in mutant proteins would lead to severe structural change in the protein by disrupting the catalytic domain or influencing the oligomerization process (**Fig. S10A,B**). Our analysis identifies previously unknown cysteine forming mutations in *DNMT3A* in a cohort of patients, which we predict would lead to loss of enzymatic function.

Somatic loss-of-Y chromosome

Somatic loss of chromosome Y (LOY) is a known marker for clonal hematopoiesis.[23] To evaluate LOY from blood cell-derived whole-exome sequencing data we measured relative sequencing coverage over the Y chromosome. Aligned sequencing reads are assigned mapping quality equal to 0 by BWA ALN[1] when an alternative equally good alignment was identified by the aligner. Such reads on the sex chromosomes paralogous regions (PAR) have less predictive value to estimate LOY as they might come from the X chromosome even when aligned to the Y chromosome. We therefore measured for each subject:

1) number of sequencing reads over the Y chromosome with mapping quality greater than 0
2) number of sequencing reads over regions X:1-2699520 (GRCh37 PAR1), X:154931044-155270560 (GRCh37 PAR2), and over regions X:88456802-92375509 and Y:2917959-6616600 (GRCh37 PAR3) with mapping quality equal to 0

We then computed the relative amount of sequencing reads for each subject by dividing those number by the total number of aligned reads over the GRCh37 human genome reference for each subject (**Fig. S12**). Although measurements were quite noisy, likely due to differences in library preparations and sequencing across samples, we could still observe that male subjects with CH-UD had overall less relative coverage over the Y-chromosome than male subjects without clonal hematopoiesis (P<0.001, Mann-Whitney test) and than male subjects with CH-CD (P=0.0089, Mann-Whitney test). Therefore LOY is either a candidate driver mutation itself, possibly due to the presence of a tumor suppressor gene in the Y chromosome, or some other event itself leading to clonal hematopoiesis is a risk factor for LOY. Interestingly, although not statistically significant, coverage for three CH-UD female subjects was also depleted over the sex chromosomes paralogous regions, possibly indicating a loss of chromosome X, an event previously observed in old women.[24]

Relationship between schizophrenia, smoking, and clonal hematopoiesis

Using a linear regression model for clonal hematopoiesis using age, sex, and schizophrenia as covariates further showed that subjects with schizophrenia had increased risk for clonal hematopoiesis (OR=1.3; 95% CI 1.1 to 1.6; P=0.0066). However, it is known that people with schizophrenia are significantly more likely to smoke and smoking is a known risk factor for several hematologic malignancies.[25–27] For less than half of the subjects with age information, 2,738 controls and 1,938 subjects with schizophrenia, we had self reported information related to whether they smoked and whether they smoked more in the past. As expected, schizophrenia was a strong risk factor for smoking (OR=7.6; 95% CI 6.2-9.3). Once smoking was included in the linear regression model, the association with schizophrenia was not observed anymore (OR=1.0; 95% CI 0.71 to 1.4). Using a linear regression model for clonal hematopoiesis using age, sex, and smoking status for subjects that stated that they either smoke and smoked more in the past or they never smoked, a statistically significant association with smoking emerged (OR=2.2; 95% CI 1.4 to 3.4; P<0.001). However, the importance of this observation needs to be further investigated, as many other variables inaccessible to us such as, for example, drinking and blood pressure, might mediate this association.

High coverage whole-genome sequencing of blood cell-derived DNA

High coverage whole-genome sequencing data were generated for Subject #1 and Subject #2 who were diagnosed with a myeloid malignancy two months after DNA sampling. Sequencing data were generated using four lanes from an Illumina HiSeq X Ten instrument for each subject with pair ended sequencing reads of 151 base pairs each and aligned against the GRCh37 human genome reference using BWA MEM version 0.7.7.[11] Base pairs across the genome were sequenced on average 108 times per subject. Genotypes and allelic counts were computed across the genome using the Haplotype Caller from the Genome Analysis Toolkit version 3.2-2. Mutations of interest were further filtered out if:
1) already in the 1000 Genomes Project phase 1 dataset[9] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/ALL.wgs.integrated_phase1_v3.20101123.snps_indels_sv.sites.vcf.gz)
2) excluded from high confidence regions for the Genome in a Bottle genotype calls for NA12878[28] (ftp://ftp.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/union13callableMQonlymerged_addcert_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_v2.18_2mindatasets_5minYesNoRatio.bed.gz)
3) excluded from the strict mask of the 1000 Genomes Project phase 1[9] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/20120824_strict_mask.bed)
4) within low complexity regions[5] (https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs37d5.bed.gz)
6) present in more than two percent of the reads from each subject
These filters defined a dataset of 69,104 mutations across ~50% of the GRCh37 human genome reference and ~60% of the coding regions. When looking at mutations that failed the hypothesis that

the alternate allelic count was distributed as a binomial process with mean 0.5 with a designed false positive rate of 0.01 or mutations at loci sequenced on average more than 200 times per subject, we observed that several of these mutations were clustering in hotspots. Upon further inspection, most of these calls were due to misalignment due to a paralogous region that was partially deleted in the human genome reference. We therefore further filtered out these mutations whenever they were found to be less than 1,000bp from each other, further defining a refined dataset of 67,919 mutations across the two subjects.

This refined set of mutations had a median allelic fraction of 49% and was consistent for each subject with two clusters of mutations, one of rare inherited mutations, and one of somatic mutations from one or more hematopoietic clones. We identified 1,153 putative somatic mutations in Subject #1 and 660 putative somatic mutations in Subject #2 failing the hypothesis that the alternate allelic count was distributed as a binomial process with mean 0.5 and with a designed false positive rate of 10^{-5}, overall consistent with previously estimated numbers.[29–34] In whole-exome sequencing data we had previously observed, respectively, 13 and 3 putative somatic mutations, consistent with the larger amount of somatic mutations observed in the first subject in whole-genome sequencing data. This observation was overall consistent with either the clone from Subject #1 being at higher frequency than the clone from Subject #2, or having multiple sub-clones, or having a clone which accumulated more mutations at the time of DNA sampling. All putative somatic mutations were confirmed in whole-genome sequencing data.

Whole-exome and whole-genome sequencing of bone marrow biopsies

Whole-exome sequencing data and low coverage whole-genome sequencing data of bone marrow biopsies were generated for Subject #2 and Subject #3. DNA was obtained from the diagnostic specimen available at the Clinical Genetics Department at Uppsala University (biobank application Bba-827-2014-064). 85 ng/µl and 88 ng/µl were obtained for, respectively, Subject #2 and Subject #3 in 10 µl water. The ThruPLEX-FD kit (Rubicon Genonics) was used to prepare three separate sequencing libraries from each subject starting from 2 µl of DNA. The three libraries were then pooled and subjected to exome capture using the SeqCap EZ Human Exome Library v3.0 kit according to standard protocols. Additionally, a fourth library was prepared with a separate index to perform low-pass whole-genome sequencing to assess the karyotypic profile of each subject. The pool of the three exome captured sequencing libraries for each individual was sequenced on one third of an Illumina Rapid Run flowcell (Hiseq 2500) at the Science for Life Laboratory in Sweden. The low-pass whole-genome libraries were spiked in at a concentration of 1 % each yielding 2.9 million read-pairs for Subject #2 and 3.2 million read-pairs for Subject #3. Sequencing reads of 101 base pairs each were aligned against the GRCh37 human genome reference using BWA MEM version 0.7.7.[11] Genotypes and allelic counts were computed across the genome using the Haplotype Caller from the Genome Analysis Toolkit version 3.2-2. Mutations of interest were further filtered out if:

1) excluded from high confidence regions for the Genome in a Bottle genotype calls for NA12878[28] (ftp://ftp.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/NIST/union13callableMQonlymerged_addcert_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_noCNVs_v2.18_2mindatasets_5minYesNoRatio.bed.gz)

2) excluded from the strict mask of the 1000 Genomes Project phase 1[9] (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/accessible_genome_masks/20120824_strict_mask.bed)

3) within low complexity regions[5] (https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs37d5.bed.gz)

Subject #1

85-years old male, diagnosed with myelodysplastic syndrome 2 months after DNA sampling. Died of unspecified leukemia 15 months after first diagnosis.

Searching for mutations in genes previously observed as significantly mutated in acute myeloid leukemia[19,33] in high coverage whole-genome sequencing data at the time of DNA sampling revealed recurrent somatic mutations *ASXL1* p.G646fsX12 and *RUNX1* p.L98fsX24, as well as somatic mutations *TET2* p.Y1148fsX5, *TET2* p.N1266S, and *STAG2* p.E472_splice and further confirmed previously identified somatic mutation *SRSF2* p.P95H (**Table S9**). Mutations in *ASXL1* and *TET2* were localized in regions of low coverage or no coverage and could not be detected in whole-exome sequencing data. Mutations in *RUNX1* and *STAG2* were not called in whole-exome sequencing data because observed in, respectively, only three and two sequencing reads. The somatic mutation *ASXL1* p.G646fsX12 was at higher allelic fraction than the other candidate drivers, suggesting that this might have been the initiating lesion. Interestingly, mutations in *ASXL1* have been shown to often co-occur in myelodysplastic syndromes with mutations in genes *RUNX1* and *SRSF2*.[35] Copy number analysis of whole-genome sequencing data revealed a normal karyotype.

Subject #2

64-year-old male, diagnosed with acute leukemia 2 months after DNA sampling. Previous history unremarkable, was referred to the hematology unit due to fatigue and pancytopenia. Bone marrow examination showed a hypercellular marrow with 50% blast cells expressing CD34, CD117, CD13 and cytoplasmic MPO, i.e. AML FAB M0. Cytogenetics showed a normal karyotype. Following intense remission induction and consolidation chemotherapy, the patient obtained sustained complete remission. Four years later, he successfully underwent cystectomy due to a low differentiated urothelial cancer in the urinary bladder.

High coverage whole-genome sequencing data at the time of DNA sampling revealed a 33 base pairs somatic insertion *CEBPA* p.K313_V314ins11 in the basic leucine zipper domain of the protein and previously observed in a different subject.[36] The mutation in *CEBPA* was not called from whole-exome sequencing data due to the shorter 76 base pairs reads used. Upon further inspection of the data through the Integrative Genomics Viewer[37] we also observed a 1 base pair frameshift deletion *CEBPA* p.P70fsX90 at lower allelic fraction of ~7%, in agreement with the observation that in-frame C-terminal mutations, usually occurring in the basic-leucine zipper (bZIP) domain, are associated with frameshift N-terminal mutations in *CEBPA*.[38] This mutation was not automatically called by the Haplotype Caller from the Genome Analysis Toolkit due to low allelic counts. Copy number analysis of whole-genome sequencing data both at the time of DNA sampling and at the time of diagnosis confirmed a normal karyotype (**Fig. S15**).

Whole-exome sequencing data of the bone marrow biopsy further confirmed the presence of the two *CEBPA* mutations and of three previously identified putative somatic mutations (**Table S10**). Estimated collective allelic fractions for these three putative somatic mutations increased in frequency between DNA sampling and first diagnosis (15.5% vs. 20.5%; P=0.037, left-tailed Fisher exact test).

Subject #3

75-year-old female, diagnosed with AML 34 months after DNA sampling. SLE with mainly cutaneous manifestations since 15 years which had been treated with steroids but not chemotherapy. Referred to the hematology unit due to pancytopenia, fatigue and pulmonary infection. Bone marrow examination showed a hypercellular marrow with 86% blast cells with no maturation and expressing CD34, CD117, CD13 and cytoplasmic MPO, i.e. AML FAB M0. Cytogenetics showed a complex karyotype including monosomy 17 and 5q-. The patient received palliative treatment with hydroxyurea and died one month later due to the leukemia.

Whole-exome sequencing data at the time of DNA sampling revealed somatic mutation *TP53* p.R248Q at an estimated allelic fraction of 24%. Whole-exome sequencing data of the bone marrow biopsy confirmed this somatic mutation at a much higher estimated allelic fraction of 86% (**Table S11**). Copy number analysis from low coverage whole-genome sequencing data confirmed that the malignancy was monosomy for chromosome 17,[39] had a partial loss of chromosome arm 5q,[40] and a complex karyotype pattern involving chromosomes 12, 13, 16, and 19 (**Fig. S15**), events that tend to co-occur in myeloid malignancies with *TP53* mutations.[41]

To test if these events were already present at the time of DNA sampling, we analyzed allelic fractions for the following six regions deleted in the malignancy:

1) chromosome 17 (**Fig. S16A**)
2) chromosome arm 5q from Mbp 72 to Mbp 155 (**Fig. S16B**)
3) chromosome arm 12p up to Mbp 26 (**Fig. S16C**)
4) chromosome arm 13q from Mbp 91 (**Fig. S16D**)
5) chromosome arm 16q (**Fig. S16E**)
6) chromosome arm 19q up to Mbp 35 (**Fig. S16F**)

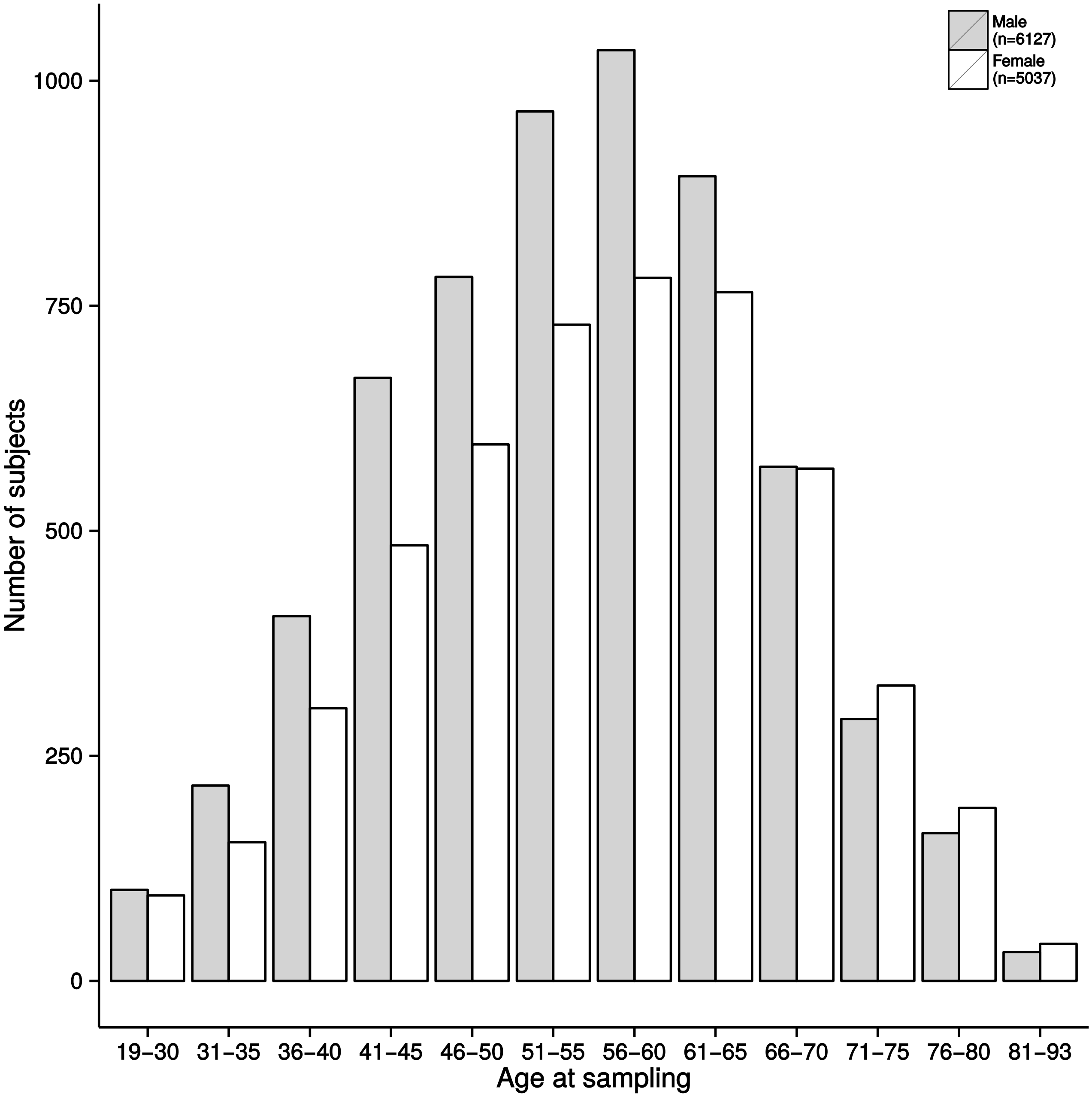
For each region we tested if allelic fractions for alleles retained in the malignancy and alleles lost in the malignancy were significantly different at the time of DNA sampling using a Mann-Whitney test. This test resulted significant for chromosome 17 (45.5% vs. 48.3%; P<0.001, **Fig. S16B**), for the chromosome arm 5q region (43.0% vs. 51.6%; P<0.001, **Fig. S16A**), but not for each of the remaining regions (**Fig. S16C–F**). High allelic fractions for these events in the biopsy shows that they needed to co-exist in the same sub-clone, this analysis suggests a most likely sequence of events of first loss of chromosome arm 5q, then loss of chromosome 17, and last the complex karyotype pattern of gains and losses on chromosomes 12, 13, 16, and 19. Therefore, while karyotyping abnormalities for chromosomes 5 and 17 must have already been present at the time of DNA sampling, 34 months before AML diagnosis, abnormalities at chromosomes 12, 13, 16, and 19 either developed later or were at undetectable frequency at the time of DNA sampling.

Statistics and figures

Cox proportional hazards analyses and Kaplan–Meier plots were performed and generated using the R **survival** package (http://cran.r-project.org/web/packages/survival/). Forest plots were generated using the R **metafor** package (http://cran.r-project.org/web/packages/metafor/). All remaining figures were generated using the R **ggplot2** package (http://cran.r-project.org/web/packages/ggplot2/) and Google Drawings (https://docs.google.com/drawings/).

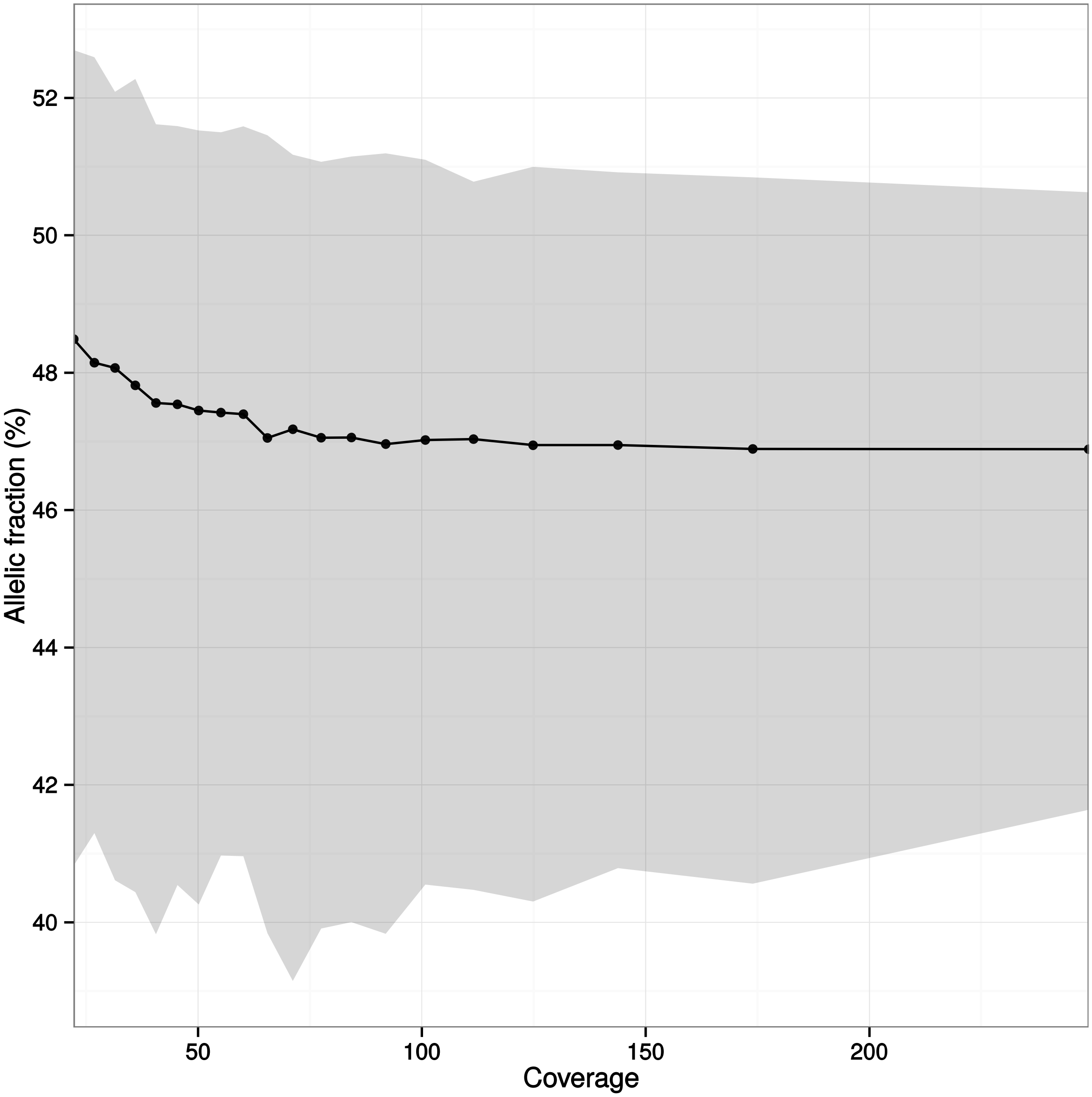
Figures

Figure S1



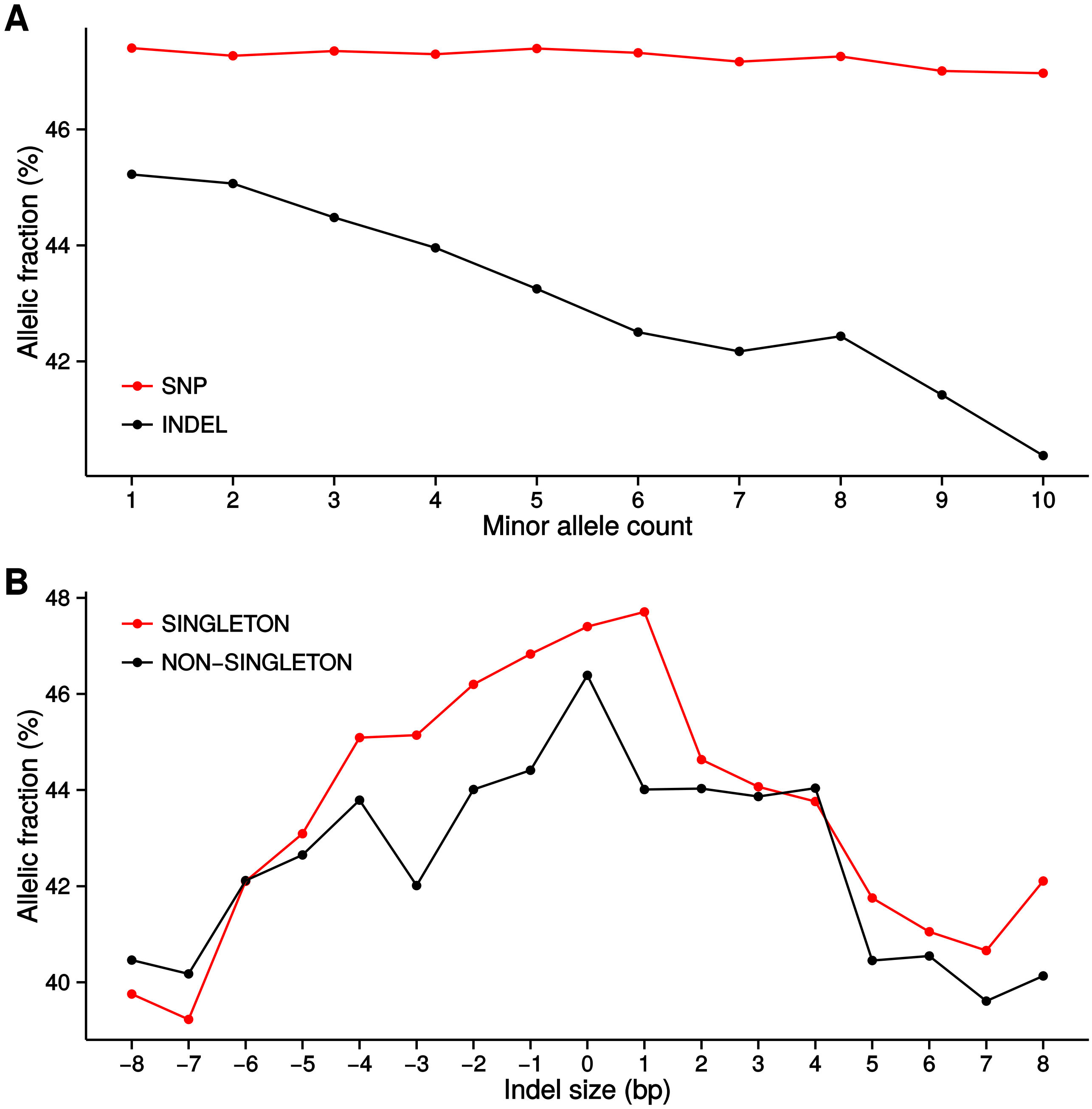
Age distribution for the 11,164 subjects for whom age at sampling information was available, stratified by sex. Because many samples were ascertained for schizophrenia and bipolar phenotypes, males in this cohort tended to be younger.

Figure S2



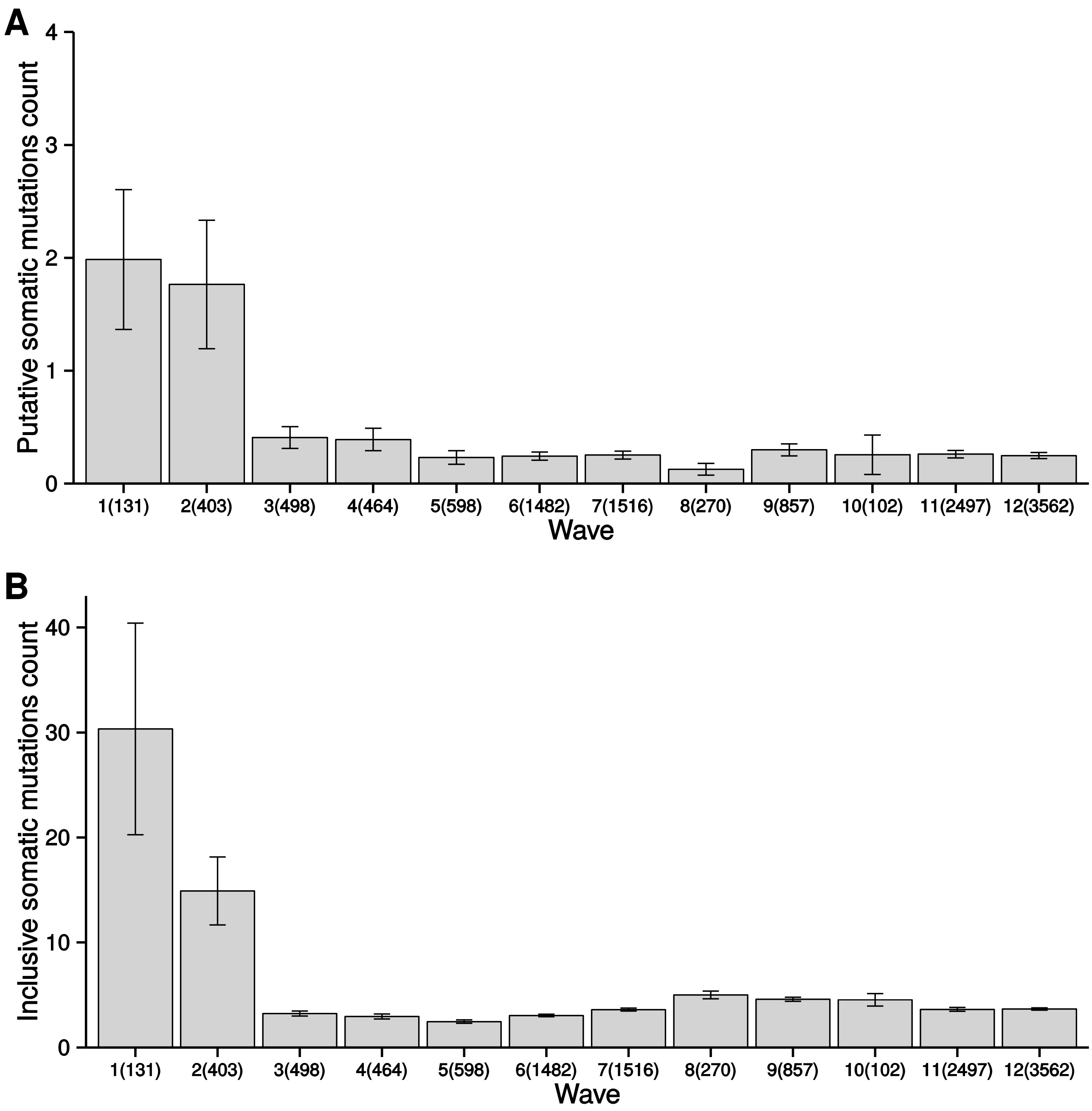
Average allelic fractions and 95% confidence interval computed for each common variant with minor allele count greater than 1000 across 12,380 subjects (minor allele frequency >4%) as a function of coverage.

Figure S3



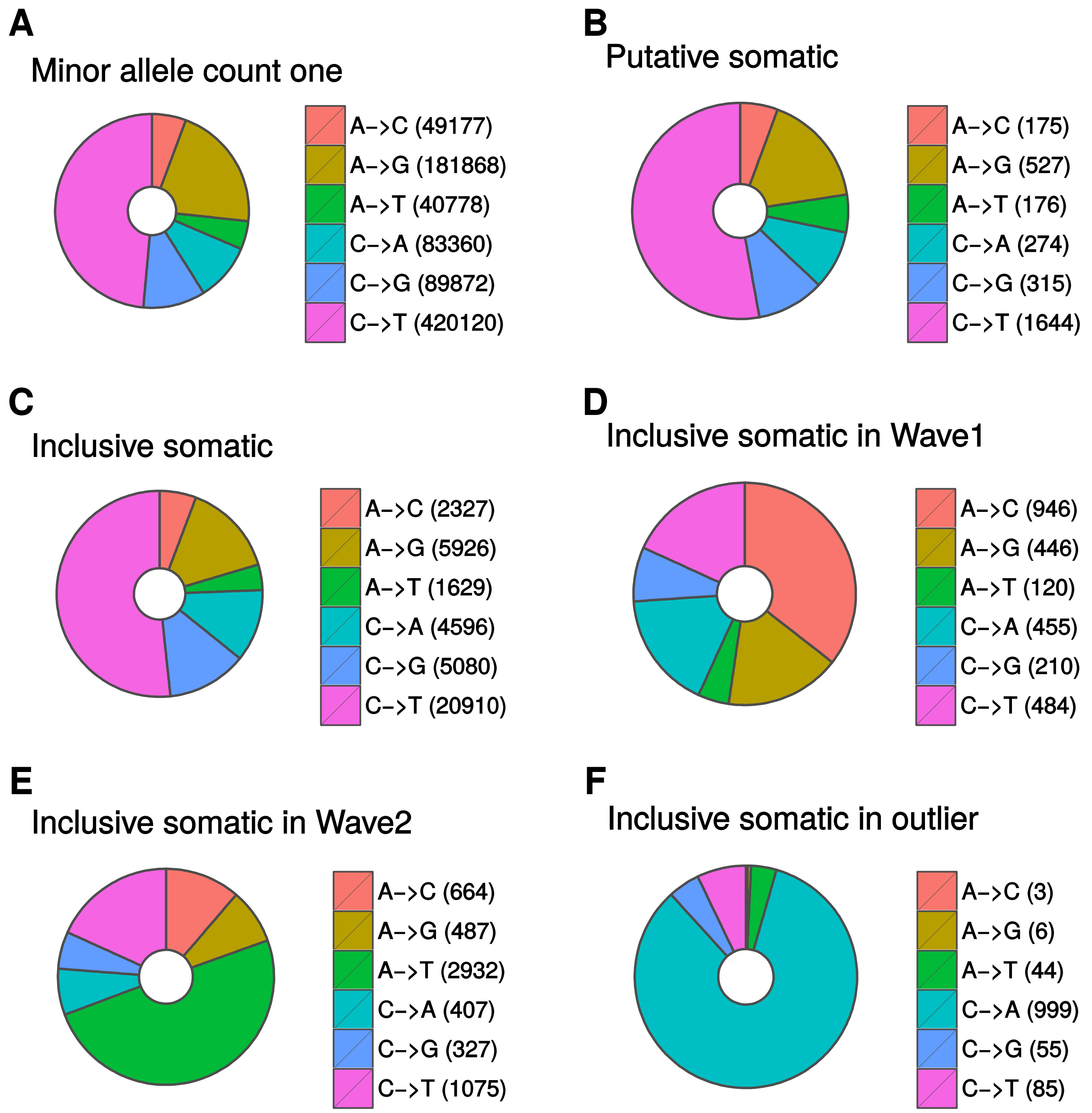
Average allelic fractions for variants with minor allele count less than 10 (minor allele frequency <0.04%) detected in the Sweden cohort using the Haplotype Caller walker from the Genome Analysis Toolkit without applying any filters. Panel A shows average allelic fractions for SNPs (in red) and indels (in black) and stratified by minor allele count. Panel B shows average allelic fractions for singletons (in red) and non-singletons (in black) alleles stratified by indel size, with positive size representing insertions and negative size representing deletions.

Figure S4



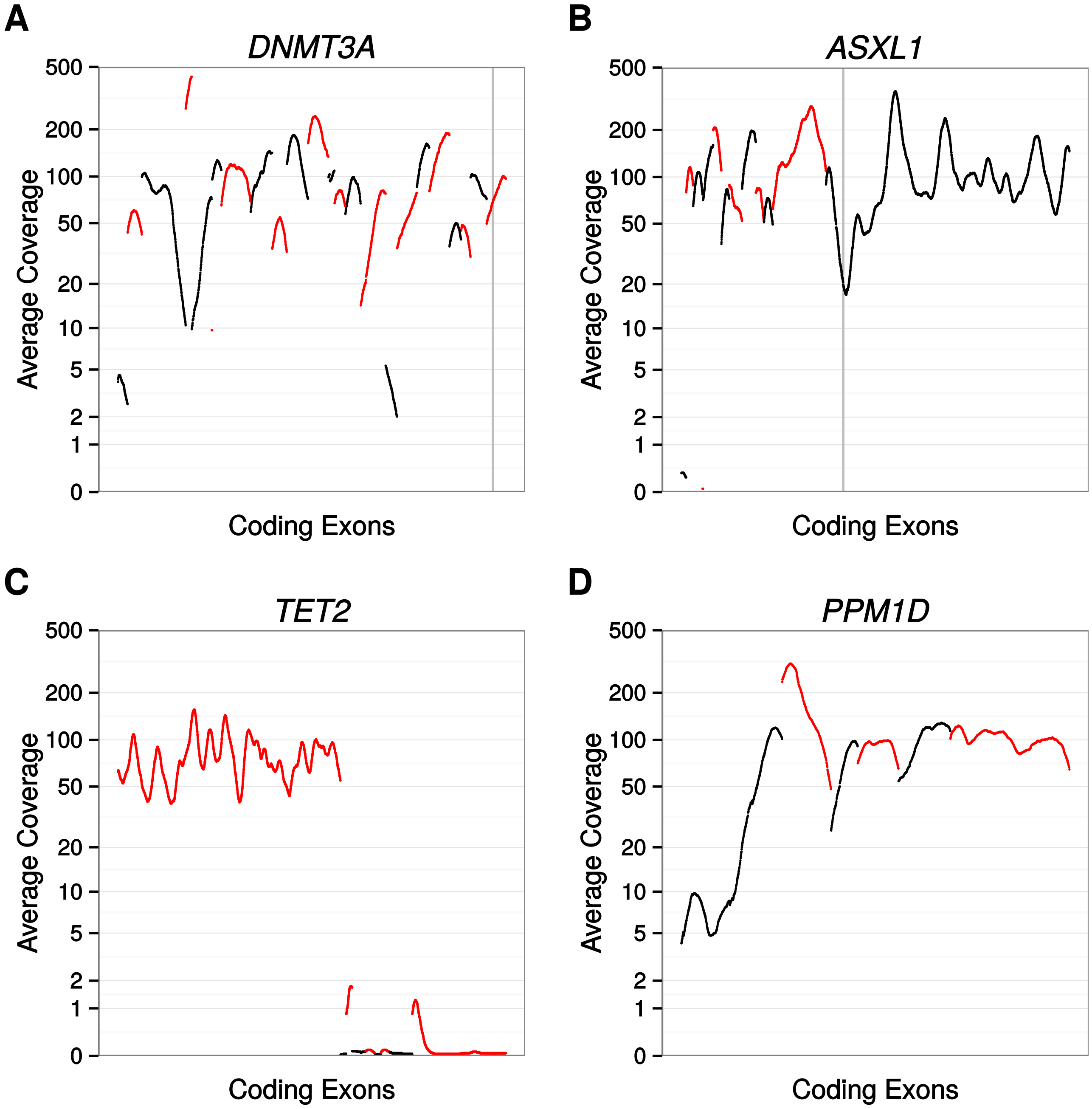
Putative somatic mutations detected across sequencing waves. Panel A and B show, respectively putative and inclusive somatic mutations stratified by sequencing waves. The first two waves exhibit an increase in detection of somatic mutations likely due to older protocols used for library preparation and sequencing.

Figure S5



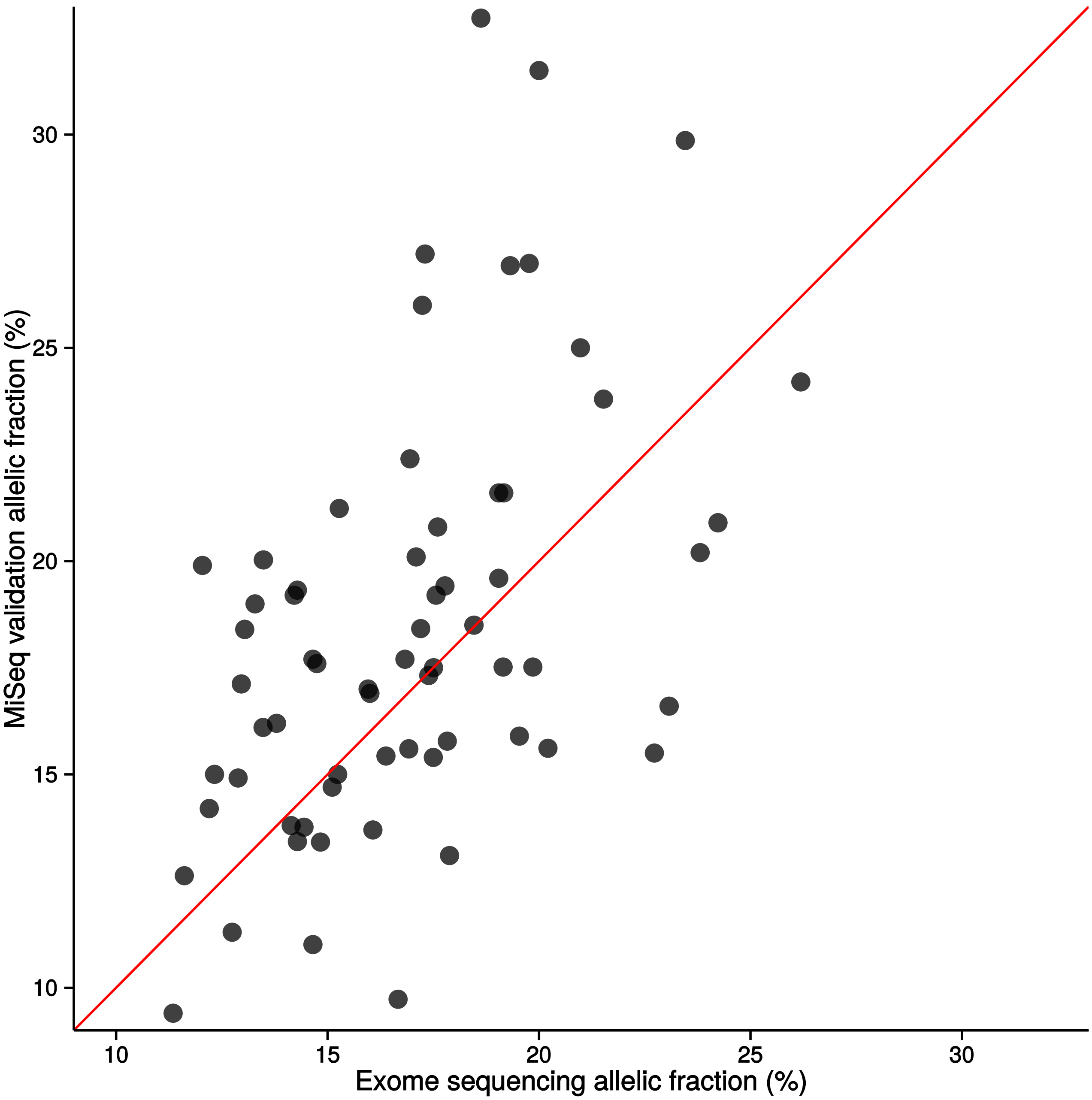
Mutation profiles for different mutation groups. Panel A shows the profile for mutations observed once or twice (minor allele frequency <0.01%) in the cohort. Panel B shows profile for putative somatic mutations from waves 3 to 12 excluding one outlier from wave 6. Panel C shows profile for inclusive somatic mutations from waves 3 to 12 excluding one outlier from wave 6. Panel D shows profile for inclusive somatic mutations from wave 1. Panel E shows profile for inclusive somatic mutations from wave 2. Panel F shows profile for inclusive somatic mutations from the outlier from wave 6.

Figure S6



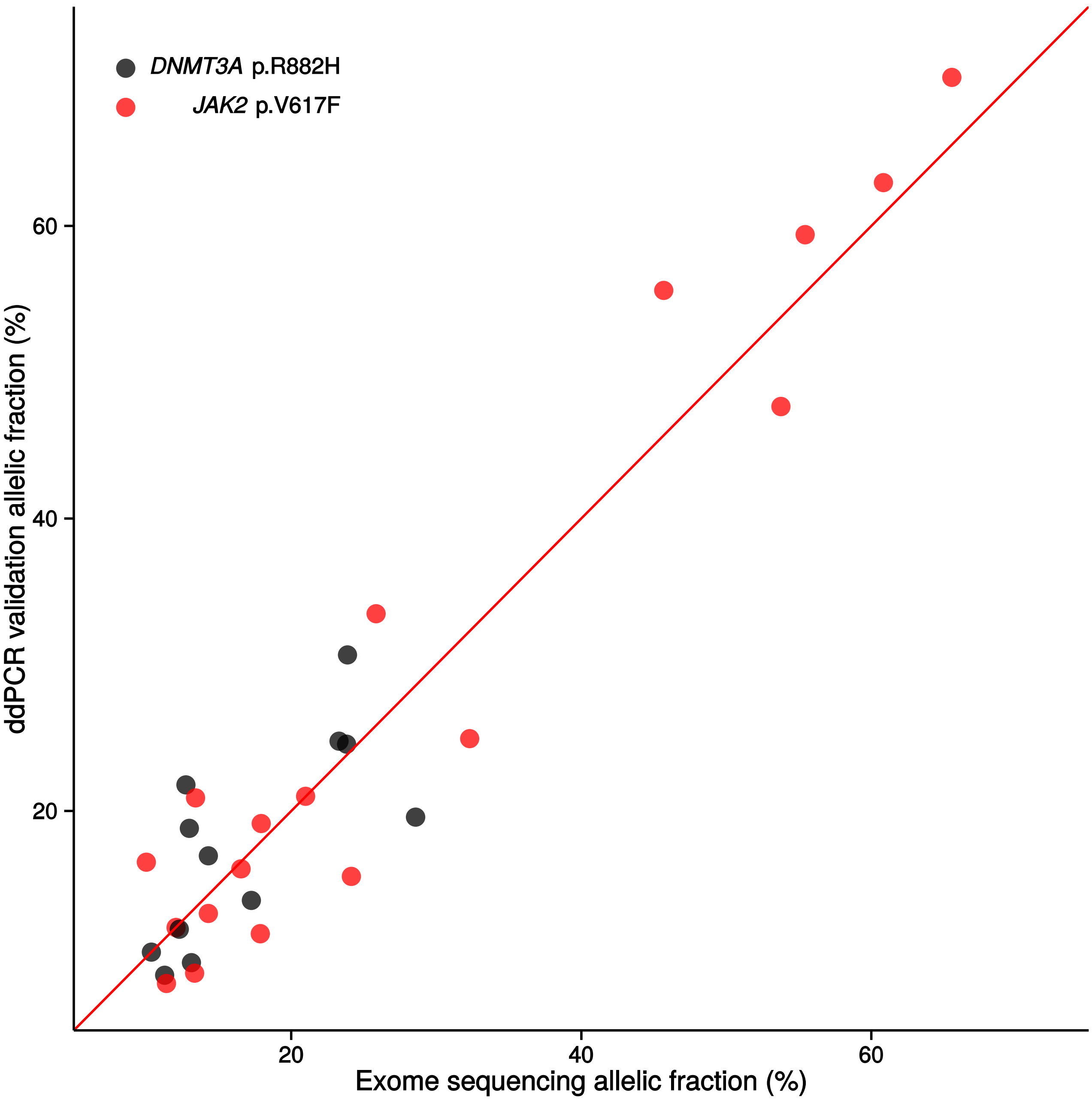
Average sequencing coverage across the coding regions for genes (A) *DNMT3A*, (B) *ASXL1*, (C) *TET2*, and (D) *PPM1D* across sequencing data from the 12,380 subjects from this study. Libraries were enriched with Agilent SureSelect Human All Exon v.2 Kit. Consecutive exons are displayed with alternating colors. Vertical gray lines show the localization of recurrent mutations *DNMT3A* p.R882H and *ASXL1* p.G646fsX12. For *DNMT3A*, exon 2 (amino acids 1–24) and exon 16 (amino acids 618–646) were sequenced on average less than 5 times per subject. The eight base-pair mononucleotide guanine nucleotide repeat giving rise to the recurrent *ASXL1* p.G646fsX12 frameshift mutation was sequenced on average less than 20 times per subject. For *TET2*, only exon 3 (amino acids 1–1166) shows coverage, most likely because only the *TET2* short isoform (NM_017628) was baited but not the *TET2* long isoform (NM_001127208).

Figure S7



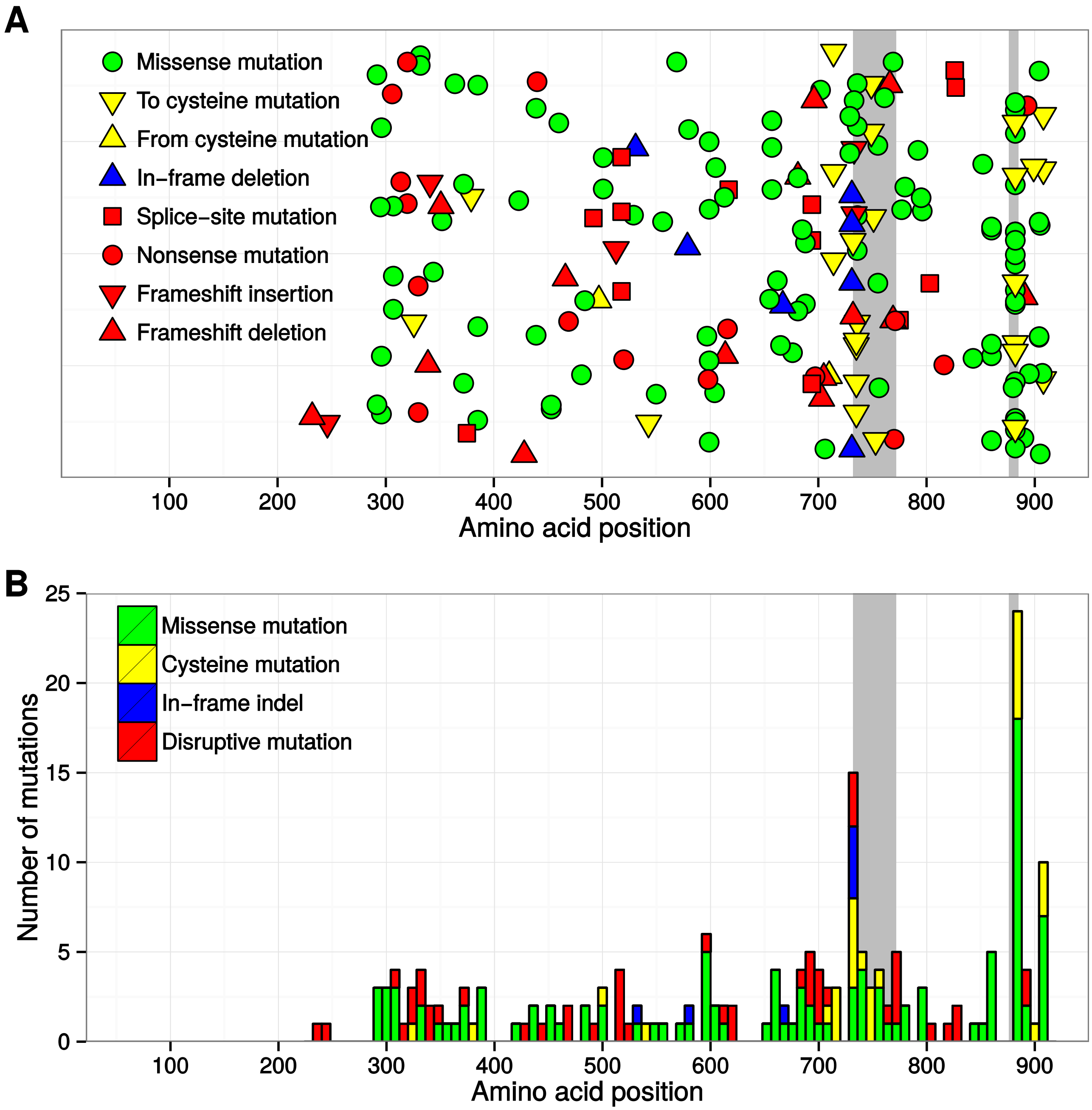
Validation experiment for 65 putative somatic mutations and candidate driver somatic mutations from 12 subjects selected for carrying one or more candidate driver somatic mutations using an Illumina MiSeq instrument. Pearson's correlation coefficient for the allelic fractions in the two experiments was r^2 0.25 (P<0.001).

Figure S8



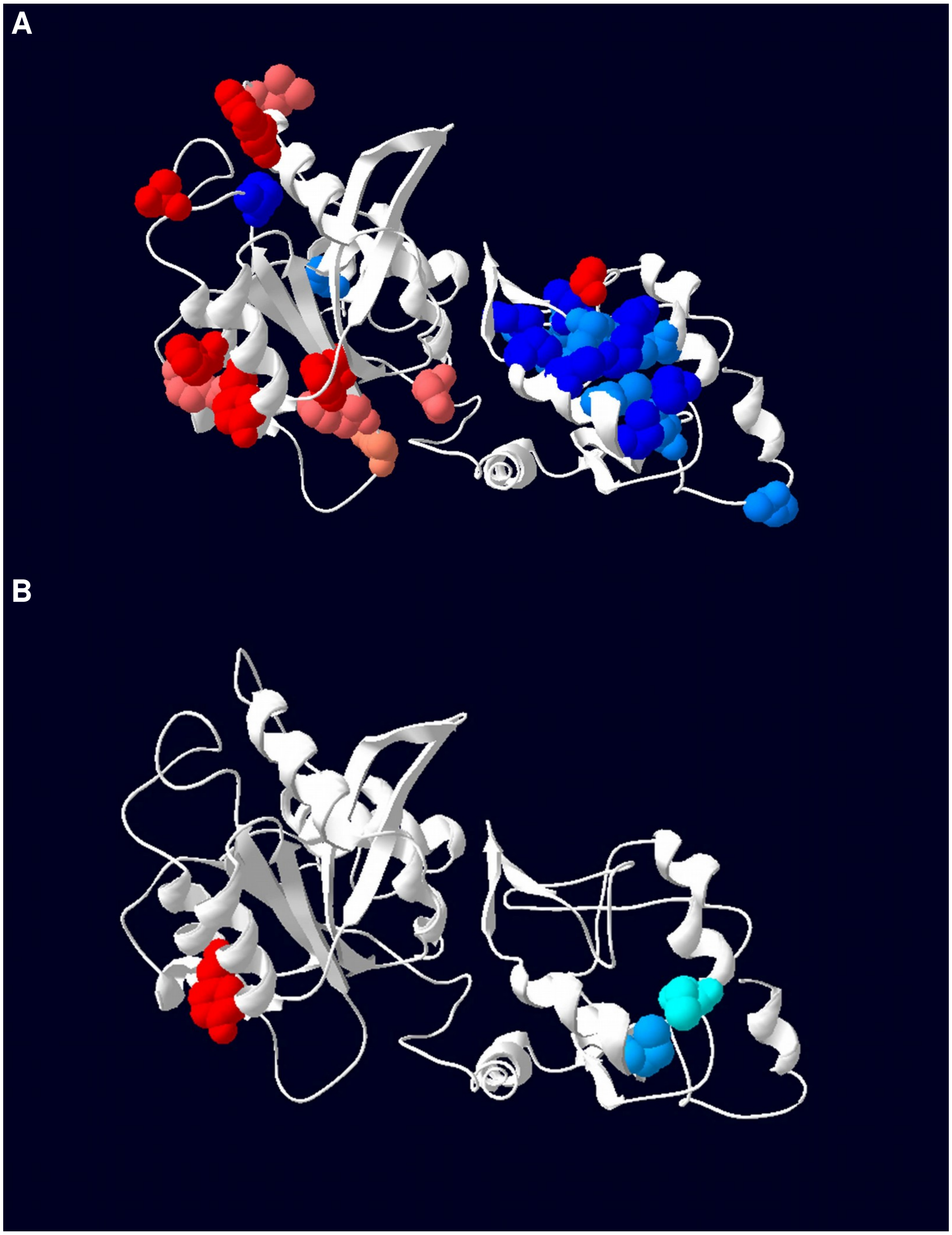
Validation experiment for 30 candidate driver somatic mutations, 18 *JAK2* p.V617F mutations, and 12 *DNMT3A* p.R882H mutations, using a droplet-based digital PCR (ddPCR) system. Pearson's correlation coefficient for the allelic fractions in the two experiments was r^2 0.90 (P<0.001).

Figure S9



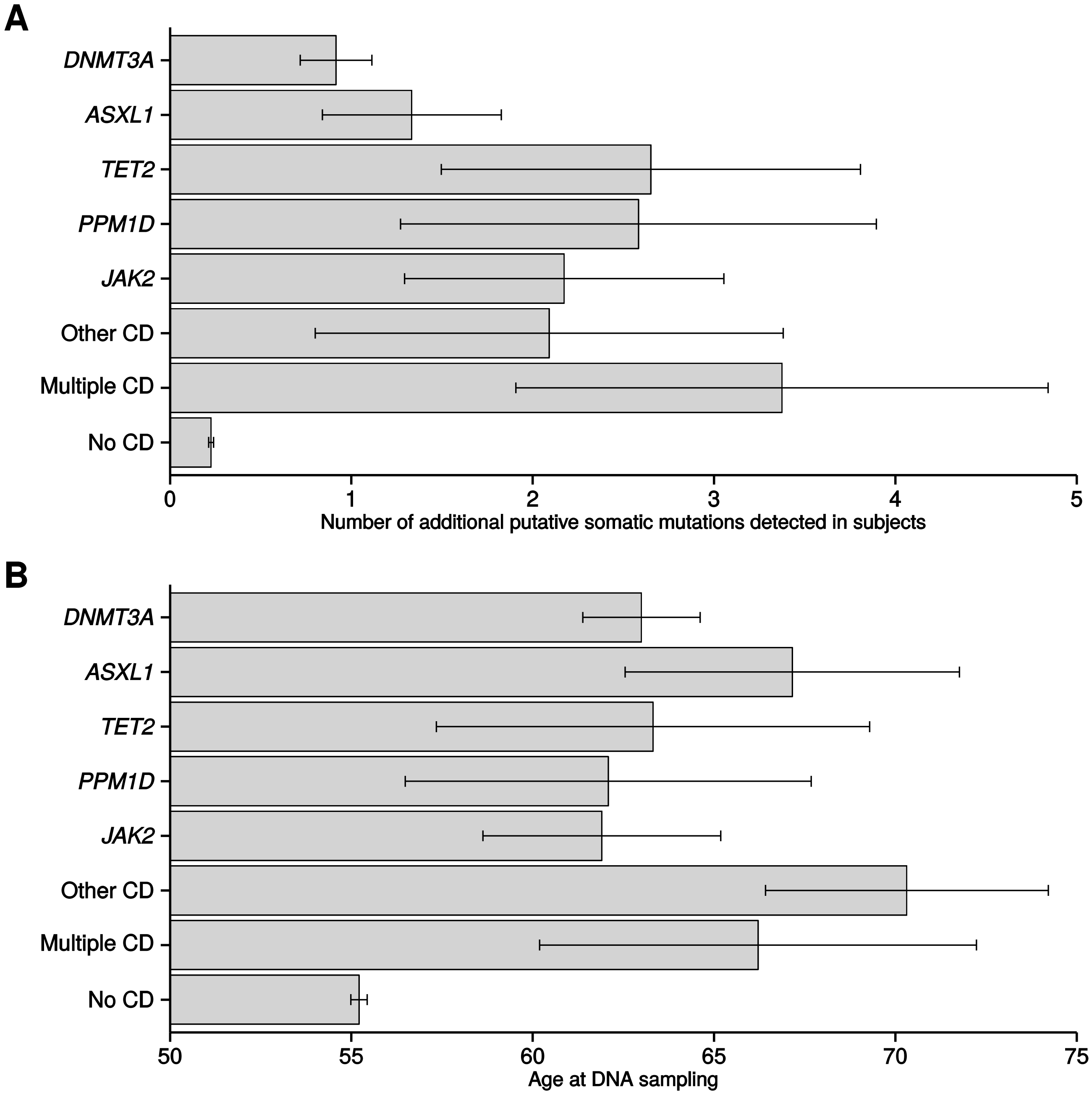
Mutations observed in the *DNMT3A* gene. Mutations across the 12,380 subjects in the cohort are visualized in Panel A as a jitterplot and in Panel B as a histogram. Amino acid regions from the FF interface (from F732 to F772) and the RD interfact (from D876 to R885) are highlighted in gray.

Figure S10



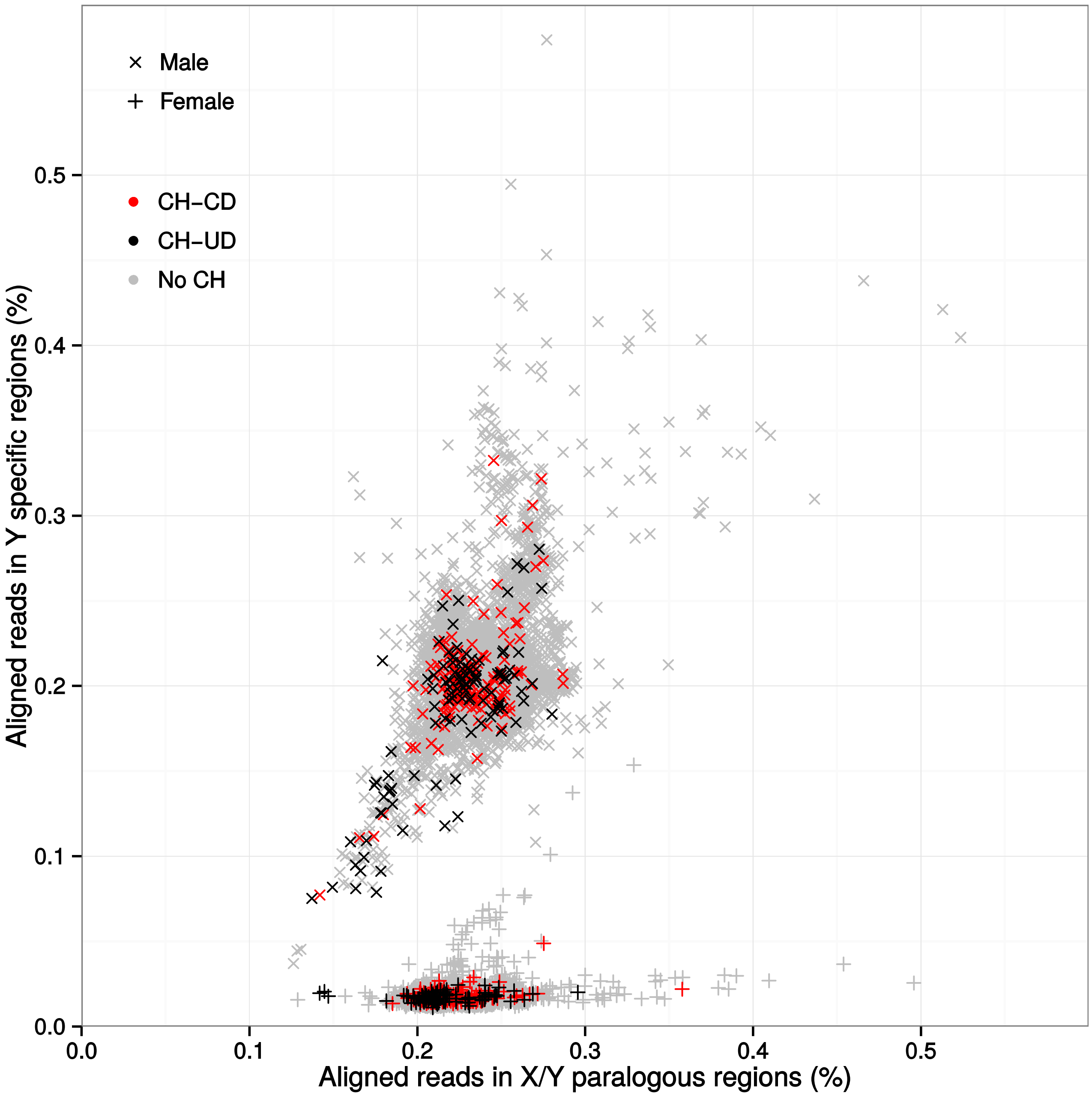
Tertiary structure for *DNMT3A* and cysteines introduced by mutations. In Panel A we show the predicted tertiary structure of *DNMT3A* (51% of the protein sequence, from R476 to F909) showing wild-type cysteine residues (in blue) and amino acid residues substituted into cysteine (in red) found in our analysis. In Panel B we show an example of a predicted disulfide bond in the mutant *DNMT3A* (F732C) using the DiANNA tool whereby the mutant C732 is predicted to form a disulfide bond with C497 (cyan). Alternatively, these de novo cysteine-forming mutations may also influence the oligomerization dynamics of *DNMT3A* due to their propensity to exist in the FF and RD domains.

Figure S11



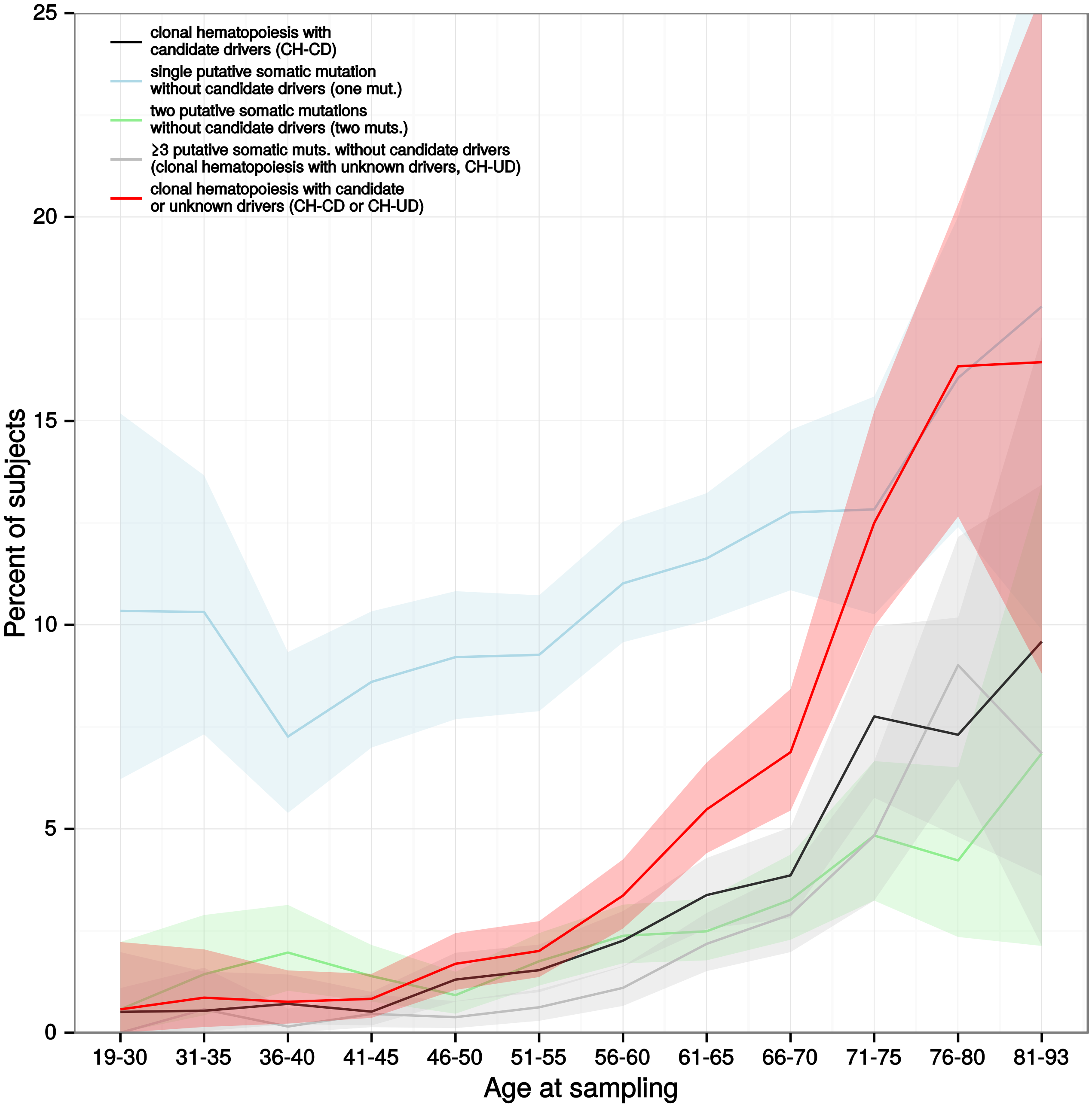
Subjects with candidate driver somatic mutations. Panel A and panel B show, respectively, average number of additional putative somatic mutations and average age for individuals carrying candidate driver somatic mutations (CD), together with 95% confidence intervals, in the most commonly mutated genes *DNMT3A*, *ASXL1*, *TET2*, *PPM1D*, *JAK2*, and other candidate driver genes grouped together. Subjects with multiple candidate driver somatic mutations or with no such mutations are separately indicated.

Figure S12



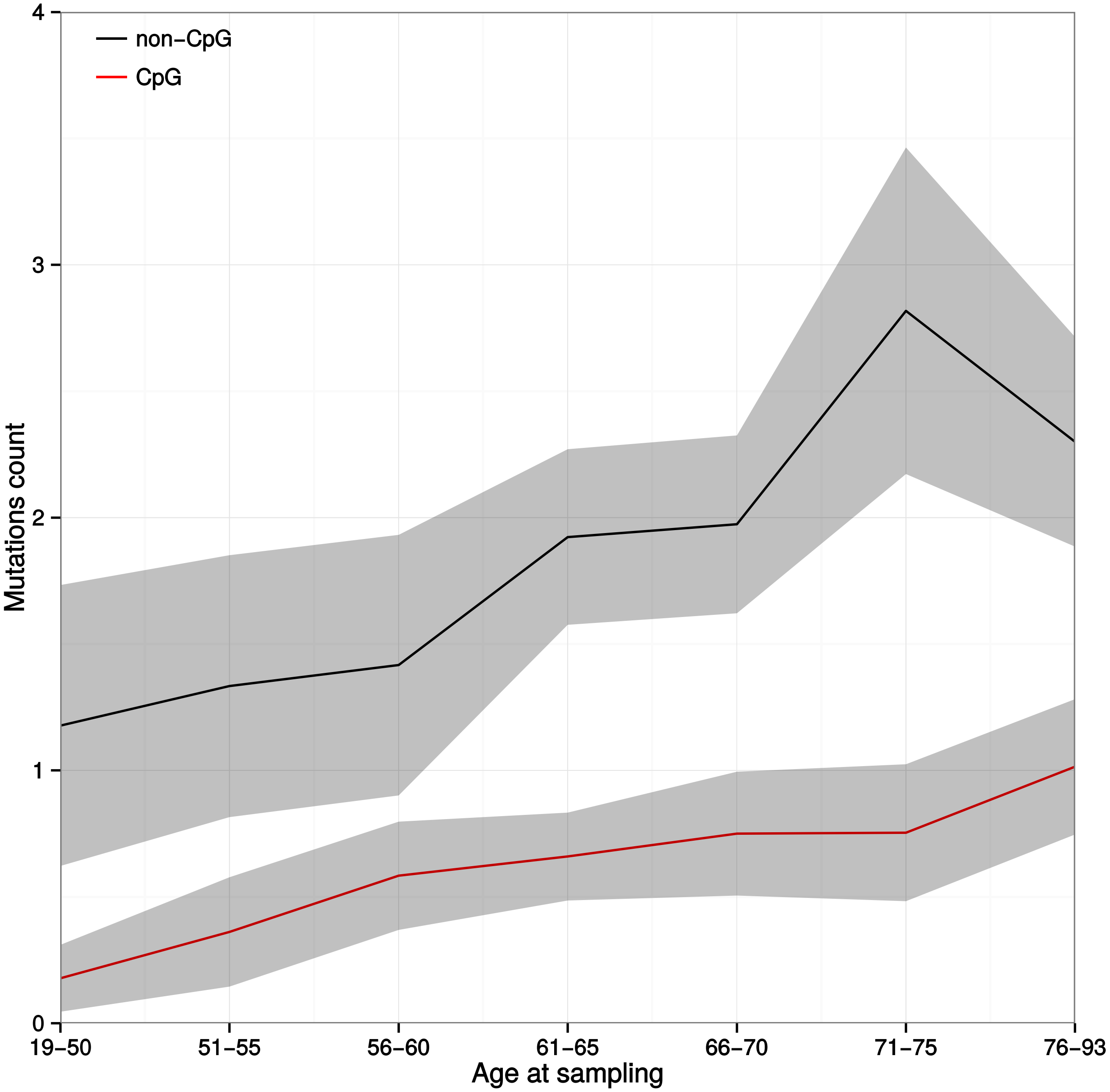
Scatterplot for sequencing reads coverage over Y chromosome. For each subject we plotted the percentage of reads aligned to the paralogous regions of the X and Y chromosomes against the percentage of reads uniquely aligned to the Y chromosome. Subjects with clonal hematopoiesis with candidate drivers (CH-CD) and with unknown drivers (CH-UD) are colored, respectively, in red and black.

Figure S13



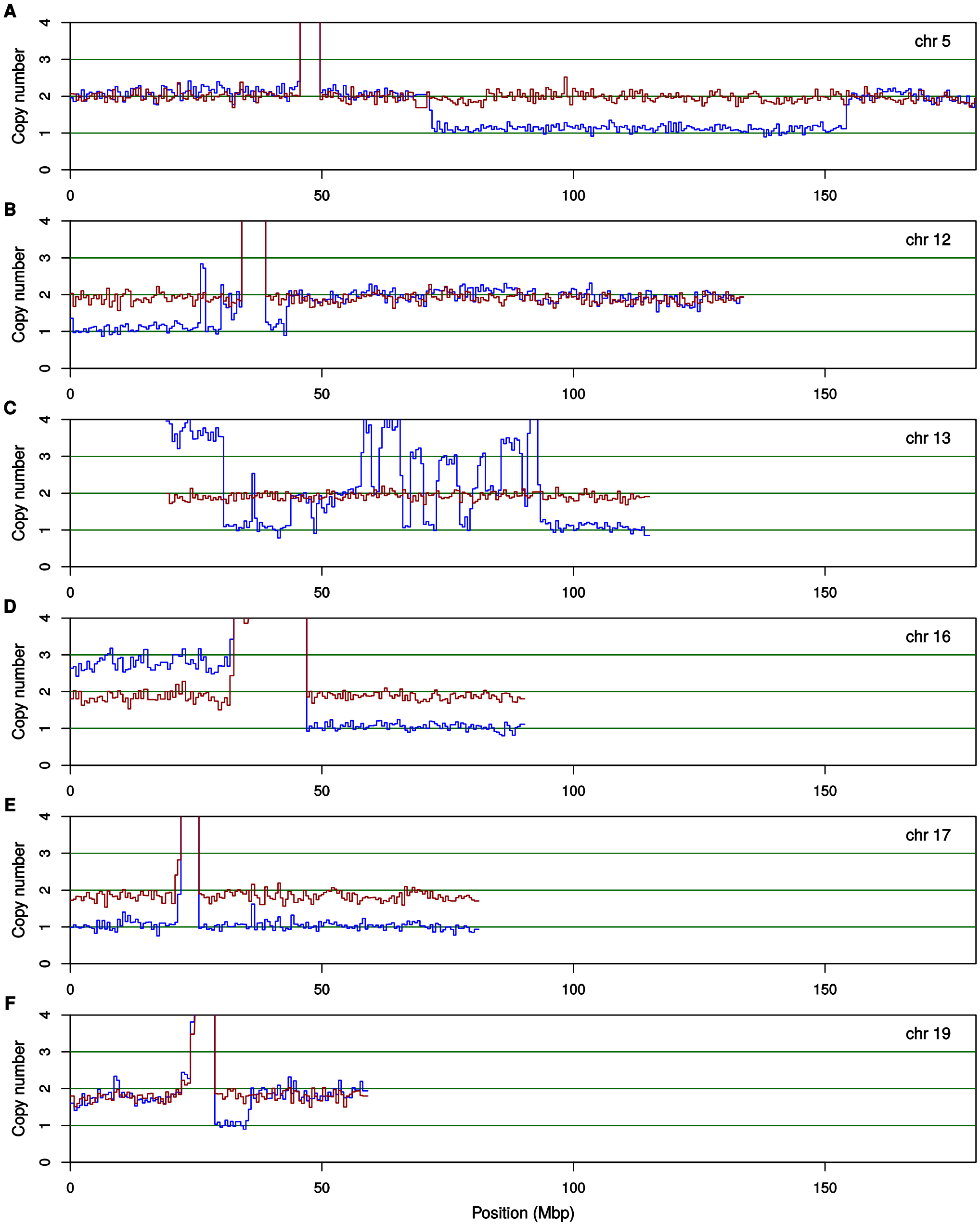
Prevalence of clonal hematopoiesis as a function of age. Percentage of subjects with clonal hematopoiesis with candidate drivers (CH-CD, in black), subjects carrying exactly one putative somatic mutation and no candidate drivers (one mut., in blue), subjects with exactly two putative somatic mutations and no candidate drivers (two muts., in green), subjects with three or more detectable somatic mutations and no candidate drivers (CH-UD, in gray), and subjects with clonal hematopoiesis with candidate or unknown drivers (CH-CD or CH-UD, in red) within 5-year age bins. Colored bands represent 95% confidence intervals.

Figure S14



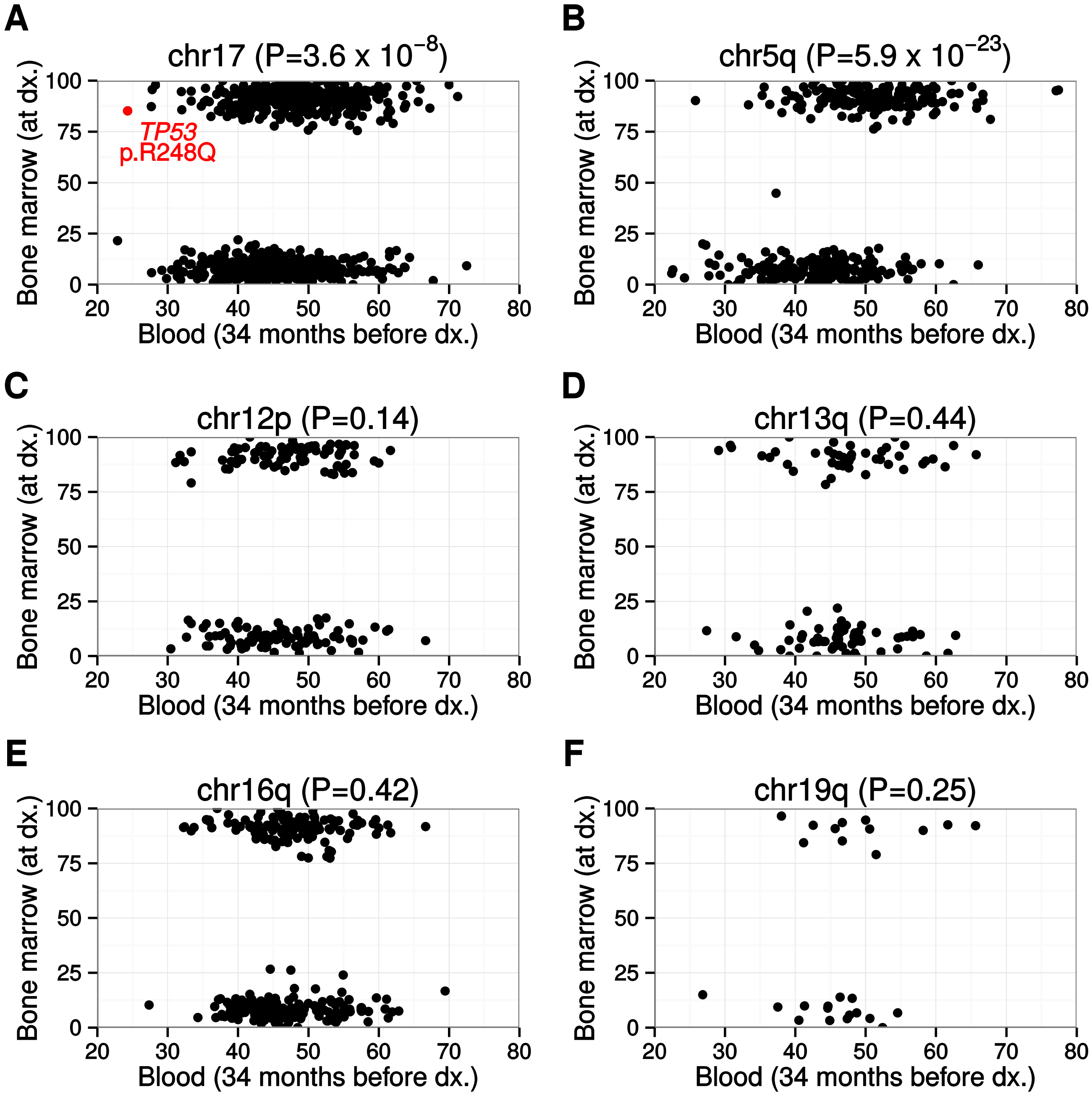
Average number of putative somatic mutations in subjects with clonal hematopoiesis as a function of age. Numbers were computed separately for non-CpG (in black) and CpG (in red) mutations within 5-year age bins. Numbers were computed for the 455 subjects with detected clonal hematopoeisis for whom age at sampling information was available. Colored bands represent 95% confidence intervals.

Figure S15



Copy number variants analysis of low coverage whole-genome sequencing data of bone marrow biopsy of Subject #2, in red, and Subject #3, in blue, at the time of first diagnosis for chromosomes 5, 12, 13, 16, 17, and 19. Copy number estimates near centromeres are overestimated due to misalignment of satellite sequence which is under-represented in the GRCh37 human genome reference. While data for Subject #2 shows a normal karyotype, Subject #3 shows loss of part of chromosome arm 5q, approximately from 5q13 to 5q33, monosomy for chromosome 17, and complex rearrangements involving chromosomes 12, 13, 16, and 19.

Figure S16



Allelic fraction analysis of alleles from Subject #3 localized on deleted regions. For each heterozygous allele, allelic fractions from whole-exome sequencing data of blood at DNA sampling and bone marrow biopsy at diagnosis are shown. Heterozygous alleles for which allelic fractions in blood are below 20% are excluded as these are enriched for sequencing or alignment artifacts. Panels A, B, C, D, E, F show heterozygous alleles from deleted regions in chromosomes, respectively, 17, 5, 12, 13, 16, and 19. P-values for comparing allelic fractions in blood between alleles retained (i.e. at more than 50% allelic fraction in bone marrow biopsy) and allelels lost (i.e. at less than 50% allelic fraction in bone marrow biopsy) using a Mann-Whitney test are reported.

Tables

Table S1

Mean age and standard deviation of different groups ascertained in the cohort.

Group	Count	Age
Total	12,380	55±12
Male	6,600	52±11
Male control	3,182	56±11
Male schizophrenia	2,964	53±11
Male bipolar	454	NA
Female	5,780	56±12
Female control	3,063	57±12
Female schizophrenia	2,006	55±12
Female bipolar	711	NA

Table S2

Mutations observed at least seven times in hematologic and lymphoid cancers in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database v69 (released June 2nd, 2014) and excluded from analysis in this study. Mutation *ASXL1* p.G646fsX12 is a genuine recurrent somatic mutation but due to low coverage at the site of the mutation it was impossible to distinguish true positives from PCR artifacts.

Variant	Amino acid change	COSMIC ID	Number of observations in hematopoietic and lymphoid cancer	Reason for exclusion
rs10521	*NOTCH1* p.D1698D	COSM33747 COSM1461158	11	Inherited mutation
rs3822214	*KIT* p.M541L	COSM28026	16	Inherited mutation
rs10663835	*CNDP1* p.L20_E21insL	COSM307404 COSM1683699	8	Inherited mutation
rs55980345	*PKD1L2* p.N236fsX26	COSM314177 COSM314178 COSM1684461 COSM1684462	7	Inherited mutation
rs139115934	*ASXL1* p.E1102D	COSM36205	15	Inherited mutation
rs146317894	*OR52D1* p.T204fsX33	COSM1683657	7	Inherited mutation
rs147836249	*TET2* p.F868L	COSM87107	7	Inherited mutation
NA	*ASXL1* p.G646fsX12	COSM34210 COSM1411076 COSM1658769	319	Potential PCR slippage error due to G homopolymer run
NA	*ASXL1* p.G645fsX58	COSM85923 COSM1180918	0	Potential PCR slippage error due to G homopolymer run
NA	*NOTCH1* p.V1578delV	COSM13047	15	Potential PCR slippage error due to CAC tandem repeat

Table S3

List of candidate driver somatic mutations detected in the cohort.

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
2	25,457,164	NA	T	C	31	28	NA	0	*DNMT3A*	NM_022552:exon23:c.A2723G:p.Y908C
2	25,457,164	NA	T	C	81	19	NA	0	*DNMT3A*	NM_022552:exon23:c.A2723G:p.Y908C
2	25,457,164	NA	T	C	94	29	NA	0	*DNMT3A*	NM_022552:exon23:c.A2723G:p.Y908C
2	25,457,168	NA	C	T	65	41	NA	0	*DNMT3A*	NM_022552:exon23:c.G2719A:p.E907K
2	25,457,173	NA	A	C	121	35	NA	0	*DNMT3A*	NM_022552:exon23:c.T2714G:p.L905R
2	25,457,173	NA	A	T	97	30	NA	0	*DNMT3A*	NM_022552:exon23:c.T2714A:p.L905Q
2	25,457,176	rs149095705	G	A	55	6	87007	6	*DNMT3A*	NM_022552:exon23:c.C2711T:p.P904L
2	25,457,176	rs149095705	G	A	65	21	87007	6	*DNMT3A*	NM_022552:exon23:c.C2711T:p.P904L
2	25,457,176	rs149095705	G	A	81	13	87007	6	*DNMT3A*	NM_022552:exon23:c.C2711T:p.P904L
2	25,457,176	rs149095705	G	A	88	11	87007	6	*DNMT3A*	NM_022552:exon23:c.C2711T:p.P904L
2	25,457,192	NA	G	A	67	40	NA	0	*DNMT3A*	NM_022552:exon23:c.C2695T:p.R899C
2	25,457,204	NA	C	T	82	23	335620 335621	0	*DNMT3A*	NM_022552:exon23:c.G2683A:p.V895M
2	25,457,209	NA	C	T	72	25	NA	0	*DNMT3A*	NM_022552:exon23:c.G2678A:p.W893X
2	25,457,215	NA	CG	C	51	12	NA	0	*DNMT3A*	NM_022552:exon23:c.2671_2672G
2	25,457,218	NA	C	T	59	13	1482984 256042	1	*DNMT3A*	NM_022552:exon23:c.G2669A:p.G890D
2	25,457,242	rs147001633	C	G	50	5	3356083 99740	14	*DNMT3A*	NM_022552:exon23:c.G2645C:p.R882P
2	25,457,242	rs147001633	C	G	75	9	3356083 99740	14	*DNMT3A*	NM_022552:exon23:c.G2645C:p.R882P
2	25,457,242	rs147001633	C	T	27	3	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	30	12	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	44	5	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	45	5	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	47	7	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	48	10	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	48	15	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	48	7	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	50	7	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	51	16	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	52	6	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	53	8	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	56	17	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	60	10	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,242	rs147001633	C	T	63	8	442676 52944	392	*DNMT3A*	NM_022552:exon23:c.G2645A:p.R882H
2	25,457,243	rs377577594	G	A	29	3	1166704 53042	164	*DNMT3A*	NM_022552:exon23:c.C2644T:p.R882C
2	25,457,243	rs377577594	G	A	29	8	1166704 53042	164	*DNMT3A*	NM_022552:exon23:c.C2644T:p.R882C
2	25,457,243	rs377577594	G	A	31	4	1166704 53042	164	*DNMT3A*	NM_022552:exon23:c.C2644T:p.R882C

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
2	25,457,243	rs377577594	G	A	59	10	1166704 53042	164	*DNMT3A*	NM_022552:exon23:c.C2644T:p.R882C
2	25,457,243	rs377577594	G	A	69	8	1166704 53042	164	*DNMT3A*	NM_022552:exon23:c.C2644T:p.R882C
2	25,457,243	rs377577594	G	A	77	24	1166704 53042	164	*DNMT3A*	NM_022552:exon23:c.C2644T:p.R882C
2	25,457,249	NA	T	C	58	19	120499	3	*DNMT3A*	NM_022552:exon23:c.A2638G:p.M880V
2	25,458,595	rs373014701	A	G	38	11	231568	2	*DNMT3A*	NM_022552:exon22:c.T2578C:p.W860R
2	25,458,595	rs373014701	A	G	43	12	231568	2	*DNMT3A*	NM_022552:exon22:c.T2578C:p.W860R
2	25,458,595	rs373014701	A	G	50	14	231568	2	*DNMT3A*	NM_022552:exon22:c.T2578C:p.W860R
2	25,458,595	rs373014701	A	G	86	17	231568	2	*DNMT3A*	NM_022552:exon22:c.T2578C:p.W860R
2	25,458,595	rs373014701	A	G	87	11	231568	2	*DNMT3A*	NM_022552:exon22:c.T2578C:p.W860R
2	25,458,619	NA	T	C	49	23	NA	0	*DNMT3A*	NM_022552:exon22:c.A2554G:p.M852V
2	25,458,646	NA	C	T	93	20	NA	0	*DNMT3A*	NM_022552:exon22:c.G2527A:p.G843S
2	25,458,696	NA	T	C	40	16	NA	0	*DNMT3A*	NM_022552:exon23:c.2479-2A>G
2	25,459,804	NA	C	A	28	6	NA	0	*DNMT3A*	NM_022552:exon22:c.2478+1G>T
2	25,459,837	NA	G	A	28	7	99739	1	*DNMT3A*	NM_022552:exon21:c.C2446T:p.Q816X
2	25,461,998	NA	C	T	23	5	NA	0	*DNMT3A*	NM_022552:exon21:c.2408+1G>A
2	25,462,020	NA	C	A	38	12	NA	0	*DNMT3A*	NM_022552:exon20:c.G2387T:p.G796V
2	25,462,024	NA	A	G	37	10	NA	0	*DNMT3A*	NM_022552:exon20:c.T2383C:p.W795R
2	25,462,032	NA	C	T	36	7	720761 720762	4	*DNMT3A*	NM_022552:exon20:c.G2375A:p.R792H
2	25,462,068	rs370751539	A	G	33	7	1583121	1	*DNMT3A*	NM_022552:exon20:c.T2339C:p.I780T
2	25,462,077	NA	G	C	19	13	NA	0	*DNMT3A*	NM_022552:exon20:c.C2330G:p.P777R
2	25,462,085	NA	C	T	22	11	NA	0	*DNMT3A*	NM_022552:exon21:c.2323-1G>A
2	25,463,174	NA	GAGAAATCG CGAGAT	G	152	19	NA	0	*DNMT3A*	NM_022552:exon19:c.2305_2319C
2	25,463,182	NA	G	A	144	23	231563	4	*DNMT3A*	NM_022552:exon19:c.C2311T:p.R771X
2	25,463,184	NA	G	T	169	36	1583106	1	*DNMT3A*	NM_022552:exon19:c.C2309A:p.S770X
2	25,463,187	NA	A	G	183	45	NA	0	*DNMT3A*	NM_022552:exon19:c.T2306C:p.I769T
2	25,463,195	NA	CTT	C	61	33	NA	0	*DNMT3A*	NM_022552:exon19:c.2296_2298G
2	25,463,212	NA	T	C	84	89	NA	0	*DNMT3A*	NM_022552:exon19:c.A2281G:p.M761V
2	25,463,225	NA	C	A	173	42	NA	0	*DNMT3A*	NM_022552:exon19:c.G2268T:p.E756D
2	25,463,229	NA	A	G	126	20	NA	0	*DNMT3A*	NM_022552:exon19:c.T2264C:p.F755S
2	25,463,229	NA	A	G	43	16	NA	0	*DNMT3A*	NM_022552:exon19:c.T2264C:p.F755S
2	25,463,234	NA	C	G	105	30	NA	0	*DNMT3A*	NM_022552:exon19:c.G2259C:p.W753C
2	25,463,241	NA	A	C	193	31	NA	0	*DNMT3A*	NM_022552:exon19:c.T2252G:p.F751C
2	25,463,248	NA	G	A	153	47	219133	4	*DNMT3A*	NM_022552:exon19:c.C2245T:p.R749C
2	25,463,248	NA	G	A	90	23	219133	4	*DNMT3A*	NM_022552:exon19:c.C2245T:p.R749C
2	25,463,286	rs139293773	C	T	137	25	1318940 133737	6	*DNMT3A*	NM_022552:exon19:c.G2207A:p.R736H
2	25,463,286	rs139293773	C	T	44	36	1318940 133737	6	*DNMT3A*	NM_022552:exon19:c.G2207A:p.R736H
2	25,463,286	rs139293773	C	T	55	32	1318940 133737	6	*DNMT3A*	NM_022552:exon19:c.G2207A:p.R736H
2	25,463,286	rs139293773	C	T	84	12	1318940 133737	6	*DNMT3A*	NM_022552:exon19:c.G2207A:p.R736H
2	25,463,287	NA	G	A	71	18	231560	5	*DNMT3A*	NM_022552:exon19:c.C2206T:p.R736C
2	25,463,289	rs147828672	T	C	100	25	133126	4	*DNMT3A*	NM_022552:exon19:c.A2204G:p.Y735C
2	25,463,289	rs147828672	T	C	76	21	133126	4	*DNMT3A*	NM_022552:exon19:c.A2204G:p.Y735C
2	25,463,289	rs147828672	T	C	84	15	133126	4	*DNMT3A*	NM_022552:exon19:c.A2204G:p.Y735C
2	25,463,289	rs147828672	T	C	90	13	133126	4	*DNMT3A*	NM_022552:exon19:c.A2204G:p.Y735C
2	25,463,295	NA	T	C	66	10	NA	0	*DNMT3A*	NM_022552:exon19:c.A2198G:p.E733G

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
2	25,463,296	NA	CAA	C	79	20	NA	0	*DNMT3A*	NM_022552:exon19:c.2195_2197G
2	25,463,296	NA	C	CA	23	3	NA	0	*DNMT3A*	NM_022552:exon19:c.2197_2197delinsTG
2	25,463,296	NA	C	CA	48	19	NA	0	*DNMT3A*	NM_022552:exon19:c.2197_2197delinsTG
2	25,463,297	NA	AAAG	A	107	26	1583117 99742	8	*DNMT3A*	NM_022552:exon19:c.2193_2196T
2	25,463,297	NA	AAAG	A	138	35	1583117 99742	8	*DNMT3A*	NM_022552:exon19:c.2193_2196T
2	25,463,297	NA	AAAG	A	77	20	1583117 99742	8	*DNMT3A*	NM_022552:exon19:c.2193_2196T
2	25,463,297	NA	AAAG	A	92	22	1583117 99742	8	*DNMT3A*	NM_022552:exon19:c.2193_2196T
2	25,463,298	NA	A	C	101	18	NA	0	*DNMT3A*	NM_022552:exon19:c.T2195G:p.F732C
2	25,463,308	rs200018028	G	A	58	70	1318937 249142	4	*DNMT3A*	NM_022552:exon19:c.C2185T:p.R729W
2	25,463,308	rs200018028	G	A	61	22	1318937 249142	4	*DNMT3A*	NM_022552:exon19:c.C2185T:p.R729W
2	25,463,541	rs367909007	G	C	124	21	442677 87011	11	*DNMT3A*	NM_022552:exon18:c.C2141G:p.S714C
2	25,463,541	rs367909007	G	C	164	29	442677 87011	11	*DNMT3A*	NM_022552:exon18:c.C2141G:p.S714C
2	25,463,541	rs367909007	G	C	172	28	442677 87011	11	*DNMT3A*	NM_022552:exon18:c.C2141G:p.S714C
2	25,463,554	NA	A	T	79	16	249803	1	*DNMT3A*	NM_022552:exon18:c.T2128A:p.C710S
2	25,463,565	NA	C	T	117	31	NA	0	*DNMT3A*	NM_022552:exon18:c.G2117A:p.G706E
2	25,463,566	NA	CA	C	62	9	NA	0	*DNMT3A*	NM_022552:exon18:c.2115_2116G
2	25,463,574	NA	AG	A	71	24	NA	0	*DNMT3A*	NM_022552:exon18:c.2107_2108T
2	25,463,578	NA	C	T	117	18	NA	0	*DNMT3A*	NM_022552:exon18:c.G2104A:p.D702N
2	25,463,593	NA	C	A	38	24	NA	0	*DNMT3A*	NM_022552:exon18:c.G2089T:p.E697X
2	25,463,595	NA	TG	T	137	18	1583101	1	*DNMT3A*	NM_022552:exon18:c.2086_2087A
2	25,464,430	NA	C	T	33	13	NA	0	*DNMT3A*	NM_022552:exon18:c.2082+1G>A
2	25,464,430	NA	C	T	46	9	NA	0	*DNMT3A*	NM_022552:exon18:c.2082+1G>A
2	25,464,430	NA	C	T	51	8	NA	0	*DNMT3A*	NM_022552:exon18:c.2082+1G>A
2	25,464,450	rs369713081	C	T	42	5	NA	0	*DNMT3A*	NM_022552:exon17:c.G2063A:p.R688H
2	25,464,450	rs369713081	C	T	43	35	NA	0	*DNMT3A*	NM_022552:exon17:c.G2063A:p.R688H
2	25,464,459	NA	C	T	29	7	1690275 1690276	0	*DNMT3A*	NM_022552:exon17:c.G2054A:p.G685E
2	25,464,470	NA	GA	G	38	7	NA	0	*DNMT3A*	NM_022552:exon17:c.2042_2043C
2	25,464,470	NA	G	C	58	11	NA	0	*DNMT3A*	NM_022552:exon17:c.C2043G:p.I681M
2	25,464,471	NA	A	T	43	8	NA	0	*DNMT3A*	NM_022552:exon17:c.T2042A:p.I681N
2	25,464,486	NA	C	A	35	24	NA	0	*DNMT3A*	NM_022552:exon17:c.G2027T:p.R676L
2	25,464,507	NA	GAGTCCT	G	40	7	NA	0	*DNMT3A*	NM_022552:exon17:c.2000_2006C
2	25,464,520	NA	C	A	41	21	NA	0	*DNMT3A*	NM_022552:exon17:c.G1993T:p.V665L
2	25,464,529	NA	C	T	42	23	NA	0	*DNMT3A*	NM_022552:exon17:c.G1984A:p.A662T
2	25,464,544	rs368961181	C	T	17	5	NA	0	*DNMT3A*	NM_022552:exon17:c.G1969A:p.V657M
2	25,464,544	rs368961181	C	T	33	11	NA	0	*DNMT3A*	NM_022552:exon17:c.G1969A:p.V657M
2	25,464,544	rs368961181	C	T	34	10	NA	0	*DNMT3A*	NM_022552:exon17:c.G1969A:p.V657M
2	25,464,549	NA	A	T	28	7	133136	1	*DNMT3A*	NM_022552:exon17:c.T1964A:p.I655N
2	25,467,023	NA	C	A	58	9	NA	0	*DNMT3A*	NM_022552:exon16:c.1851+1G>T
2	25,467,029	NA	C	A	89	15	NA	0	*DNMT3A*	NM_022552:exon15:c.G1846T:p.E616X
2	25,467,034	NA	TC	T	81	28	NA	0	*DNMT3A*	NM_022552:exon15:c.1840_1841A
2	25,467,038	NA	G	C	44	21	NA	0	*DNMT3A*	NM_022552:exon15:c.C1837G:p.H613D
2	25,467,061	NA	A	G	62	22	NA	0	*DNMT3A*	NM_022552:exon15:c.T1814C:p.L605P
2	25,467,064	NA	C	T	40	25	NA	0	*DNMT3A*	NM_022552:exon15:c.G1811A:p.R604Q
2	25,467,078	NA	C	A	30	19	NA	0	*DNMT3A*	NM_022552:exon15:c.G1797T:p.E599D

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
2	25,467,078	NA	C	A	39	23	NA	0	*DNMT3A*	NM_022552:exon15:c.G1797T:p.E599D
2	25,467,078	NA	C	A	58	42	NA	0	*DNMT3A*	NM_022552:exon15:c.G1797T:p.E599D
2	25,467,078	NA	C	A	63	54	NA	0	*DNMT3A*	NM_022552:exon15:c.G1797T:p.E599D
2	25,467,083	NA	G	A	49	18	133736	4	*DNMT3A*	NM_022552:exon15:c.C1792T:p.R598X
2	25,467,086	NA	G	A	39	34	NA	0	*DNMT3A*	NM_022552:exon15:c.C1789T:p.R597W
2	25,467,133	NA	CAGGGGT	C	34	5	NA	0	*DNMT3A*	NM_022552:exon15:c.1736_1742G
2	25,467,136	NA	G	C	7	14	NA	0	*DNMT3A*	NM_022552:exon15:c.C1739G:p.P580R
2	25,467,169	NA	G	A	13	14	NA	0	*DNMT3A*	NM_022552:exon15:c.C1706T:p.P569L
2	25,467,410	NA	T	C	53	33	NA	0	*DNMT3A*	NM_022552:exon14:c.A1666G:p.R556G
2	25,467,428	NA	C	T	67	12	256035	4	*DNMT3A*	NM_022552:exon14:c.G1648A:p.G550R
2	25,467,449	NA	C	A	53	8	87002	10	*DNMT3A*	NM_022552:exon14:c.G1627T:p.G543C
2	25,467,481	NA	CCGT	C	37	13	1583078	1	*DNMT3A*	NM_022552:exon14:c.1592_1595G
2	25,467,490	NA	T	A	69	17	NA	0	*DNMT3A*	NM_022552:exon14:c.A1586T:p.D529V
2	25,467,516	NA	G	T	67	12	NA	0	*DNMT3A*	NM_022552:exon14:c.C1560A:p.C520X
2	25,468,120	NA	A	C	60	20	NA	0	*DNMT3A*	NM_022552:exon14:c.1554+2T>G
2	25,468,121	NA	C	T	103	12	NA	0	*DNMT3A*	NM_022552:exon14:c.1554+1G>A
2	25,468,121	NA	C	T	63	10	NA	0	*DNMT3A*	NM_022552:exon14:c.1554+1G>A
2	25,468,138	NA	A	AT	46	11	NA	0	*DNMT3A*	NM_022552:exon13:c.1538_1538delinsAT
2	25,468,174	rs149738328	T	C	37	32	231571	3	*DNMT3A*	NM_022552:exon13:c.A1502G:p.N501S
2	25,468,174	rs149738328	T	C	50	32	231571	3	*DNMT3A*	NM_022552:exon13:c.A1502G:p.N501S
2	25,468,186	NA	C	T	23	7	1318925 1318926	3	*DNMT3A*	NM_022552:exon13:c.G1490A:p.C497Y
2	25,468,888	NA	C	T	105	43	NA	0	*DNMT3A*	NM_022552:exon13:c.1474+1G>A
2	25,468,912	NA	C	T	65	1	NA	0	*DNMT3A*	NM_022552:exon12:c.G1451A:p.R484Q
2	25,468,922	NA	A	C	55	3	NA	0	*DNMT3A*	NM_022552:exon12:c.T1441G:p.Y481D
2	25,469,053	NA	C	A	125	28	NA	0	*DNMT3A*	NM_022552:exon11:c.G1405T:p.E469X
2	25,469,060	NA	CT	C	133	25	NA	0	*DNMT3A*	NM_022552:exon11:c.1397_1398G
2	25,469,080	NA	T	C	106	90	NA	0	*DNMT3A*	NM_022552:exon11:c.A1378G:p.S460G
2	25,469,100	NA	G	A	104	89	NA	0	*DNMT3A*	NM_022552:exon11:c.C1358T:p.P453L
2	25,469,100	NA	G	A	77	99	NA	0	*DNMT3A*	NM_022552:exon11:c.C1358T:p.P453L
2	25,469,139	NA	C	T	179	38	NA	0	*DNMT3A*	NM_022552:exon11:c.G1319A:p.W440X
2	25,469,142	NA	A	G	153	102	NA	0	*DNMT3A*	NM_022552:exon11:c.T1316C:p.M439T
2	25,469,142	NA	A	G	80	66	NA	0	*DNMT3A*	NM_022552:exon11:c.T1316C:p.M439T
2	25,469,174	NA	CT	C	167	24	NA	0	*DNMT3A*	NM_022552:exon11:c.1283_1284G
2	25,469,501	NA	C	G	52	70	NA	0	*DNMT3A*	NM_022552:exon10:c.G1267C:p.E423Q
2	25,469,614	NA	G	A	109	73	NA	0	*DNMT3A*	NM_022552:exon10:c.C1154T:p.P385L
2	25,469,614	NA	G	A	61	39	NA	0	*DNMT3A*	NM_022552:exon10:c.C1154T:p.P385L
2	25,469,614	NA	G	A	97	62	NA	0	*DNMT3A*	NM_022552:exon10:c.C1154T:p.P385L
2	25,469,633	NA	G	A	83	14	NA	0	*DNMT3A*	NM_022552:exon10:c.C1135T:p.R379C
2	25,469,647	NA	T	G	149	18	NA	0	*DNMT3A*	NM_022552:exon11:c.1123-2A>C
2	25,469,927	NA	A	G	23	14	NA	0	*DNMT3A*	NM_022552:exon9:c.T1115C:p.V372A
2	25,469,928	rs371677904	C	T	21	20	NA	0	*DNMT3A*	NM_022552:exon9:c.G1114A:p.V372I
2	25,469,951	NA	A	G	30	17	NA	0	*DNMT3A*	NM_022552:exon9:c.T1091C:p.M364T
2	25,469,987	rs139053291	C	T	24	11	133129	1	*DNMT3A*	NM_022552:exon9:c.G1055A:p.S352N
2	25,469,988	NA	TGC	TT	53	9	NA	0	*DNMT3A*	NM_022552:exon9:c.1052_1054AA
2	25,470,011	NA	A	G	17	6	NA	0	*DNMT3A*	NM_022552:exon9:c.T1031C:p.L344P
2	25,470,019	NA	A	AAC	23	9	NA	0	*DNMT3A*	NM_022552:exon9:c.1023_1023delinsGTT
2	25,470,028	NA	CT	C	21	6	NA	0	*DNMT3A*	NM_022552:exon10:c.1015_splice
2	25,470,479	NA	C	T	147	30	477212	0	*DNMT3A*	NM_022552:exon8:c.G995A:p.G332E

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
2	25,470,480	NA	C	T	102	48	NA	0	*DNMT3A*	NM_022552:exon8:c.G994A:p.G332R
2	25,470,484	NA	C	T	150	21	249799	1	*DNMT3A*	NM_022552:exon8:c.G990A:p.W330X
2	25,470,484	NA	C	T	72	11	249799	1	*DNMT3A*	NM_022552:exon8:c.G990A:p.W330X
2	25,470,498	NA	G	A	90	17	NA	0	*DNMT3A*	NM_022552:exon8:c.C976T:p.R326C
2	25,470,516	NA	G	A	108	17	1318922 133721 133724	4	*DNMT3A*	NM_022552:exon8:c.C958T:p.R320X
2	25,470,516	NA	G	A	98	16	1318922 133721 133724	4	*DNMT3A*	NM_022552:exon8:c.C958T:p.R320X
2	25,470,532	NA	C	T	83	30	NA	0	*DNMT3A*	NM_022552:exon8:c.G942A:p.W314X
2	25,470,554	NA	G	A	77	16	NA	0	*DNMT3A*	NM_022552:exon8:c.C920T:p.P307L
2	25,470,554	NA	G	C	51	6	221579	1	*DNMT3A*	NM_022552:exon8:c.C920G:p.P307R
2	25,470,554	NA	G	C	86	18	221579	1	*DNMT3A*	NM_022552:exon8:c.C920G:p.P307R
2	25,470,556	NA	C	T	60	10	NA	0	*DNMT3A*	NM_022552:exon8:c.G918A:p.W306X
2	25,470,588	NA	C	T	60	13	NA	0	*DNMT3A*	NM_022552:exon8:c.G886A:p.V296M
2	25,470,588	NA	C	T	83	15	NA	0	*DNMT3A*	NM_022552:exon8:c.G886A:p.V296M
2	25,470,588	NA	C	T	86	18	NA	0	*DNMT3A*	NM_022552:exon8:c.G886A:p.V296M
2	25,470,591	NA	G	C	48	10	NA	0	*DNMT3A*	NM_022552:exon8:c.C883G:p.L295V
2	25,470,599	NA	A	G	70	19	NA	0	*DNMT3A*	NM_022552:exon8:c.T875C:p.I292T
2	25,470,599	NA	A	G	99	17	NA	0	*DNMT3A*	NM_022552:exon8:c.T875C:p.I292T
2	25,471,024	NA	G	GC	71	18	NA	0	*DNMT3A*	NM_022552:exon7:c.737_737delinsGC
2	25,471,064	NA	GC	G	58	22	NA	0	*DNMT3A*	NM_022552:exon7:c.696_697C
2	198,266,834	NA	T	C	148	16	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	50	12	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	50	16	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	53	6	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	60	17	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	66	10	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	79	14	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	91	8	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,266,834	NA	T	C	97	11	84677	230	*SF3B1*	NM_012433:exon15:c.A2098G:p.K700E
2	198,267,359	rs377023736	C	A	207	27	131557	13	*SF3B1*	NM_012433:exon14:c.G1998T:p.K666N
2	198,267,359	rs377023736	C	G	66	22	132937	9	*SF3B1*	NM_012433:exon14:c.G1998C:p.K666N
2	198,267,360	NA	T	G	61	11	131556	8	*SF3B1*	NM_012433:exon14:c.A1997C:p.K666T
2	198,267,491	NA	C	G	106	15	132938	7	*SF3B1*	NM_012433:exon14:c.G1866C:p.E622D
3	38,182,641	rs387907272	T	C	91	21	85940	1027	*MYD88*	NM_002468:exon5:c.794T>C:p.L265P
4	106,155,544	NA	G	T	29	14	3428018 3428019	0	*TET2*	NM_017628:exon3:c.G445T:p.E149X
4	106,155,915	NA	GC	G	24	12	NA	0	*TET2*	NM_017628:exon3:c.816_817G
4	106,156,079	NA	C	G	97	18	NA	0	*TET2*	NM_017628:exon3:c.C980G:p.S327X
4	106,156,409	NA	A	AC	73	12	NA	0	*TET2*	NM_017628:exon3:c.1310_1310delinsAC
4	106,156,441	NA	G	T	38	9	NA	0	*TET2*	NM_017628:exon3:c.G1342T:p.E448X
4	106,156,564	NA	GA	G	106	23	NA	0	*TET2*	NM_017628:exon3:c.1465_1466G
4	106,156,623	NA	GT	G	50	11	NA	0	*TET2*	NM_017628:exon3:c.1524_1525G
4	106,156,747	NA	C	T	119	13	1318629 41644	26	*TET2*	NM_017628:exon3:c.C1648T:p.R550X
4	106,156,758	NA	G	GC	152	31	43490	3	*TET2*	NM_017628:exon3:c.1659_1659delinsGC
4	106,157,162	NA	A	AT	105	16	NA	0	*TET2*	NM_017628:exon3:c.2063_2063delinsAT
4	106,157,332	NA	CAG	C	39	28	NA	0	*TET2*	NM_017628:exon3:c.2233_2235C
4	106,157,335	NA	C	T	53	10	87099	1	*TET2*	NM_017628:exon3:c.C2236T:p.Q746X
4	106,157,367	NA	AC	A	75	39	NA	0	*TET2*	NM_017628:exon3:c.2268_2269A

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
4	106,157,467	NA	C	T	53	10	43416	1	*TET2*	NM_017628:exon3:c.C2368T:p.Q790X
4	106,157,503	NA	GT	G	66	14	NA	0	*TET2*	NM_017628:exon3:c.2404_2405G
4	106,157,525	NA	TA	T	68	11	NA	0	*TET2*	NM_017628:exon3:c.2426_2427T
4	106,157,542	NA	A	T	55	22	NA	0	*TET2*	NM_017628:exon3:c.A2443T:p.R815X
4	106,157,608	NA	AAT	A	53	20	NA	0	*TET2*	NM_017628:exon3:c.2509_2511A
4	106,157,638	NA	C	T	38	8	NA	0	*TET2*	NM_017628:exon3:c.C2539T:p.Q847X
4	106,157,761	NA	C	T	54	11	NA	0	*TET2*	NM_017628:exon3:c.C2662T:p.Q888X
4	106,157,842	NA	G	GCT	31	10	NA	0	*TET2*	NM_017628:exon3:c.2743_2743delinsGCT
4	106,158,224	NA	AC	A	97	19	NA	0	*TET2*	NM_017628:exon3:c.3125_3126A
4	106,158,349	NA	CA	C	77	12	NA	0	*TET2*	NM_017628:exon3:c.3250_3251C
4	106,158,359	NA	CTT	C	42	10	NA	0	*TET2*	NM_017628:exon3:c.3260_3262C
4	106,158,378	NA	C	CA	18	3	NA	0	*TET2*	NM_017628:exon3:c.3279_3279delinsCA
4	106,158,378	NA	C	CA	40	9	NA	0	*TET2*	NM_017628:exon3:c.3279_3279delinsCA
4	106,158,442	NA	C	CT	55	17	NA	0	*TET2*	NM_017628:exon3:c.3343_3343delinsCT
4	106,158,485	NA	AT	A	69	22	NA	0	*TET2*	NM_017628:exon3:c.3386_3387A
4	106,158,509	NA	G	A	75	24	87117	1	*TET2*	NM_001127208:exon3:c.3409+1G>A
4	106,158,579	NA	A	AT	32	23	NA	0	*TET2*	NM_017628:exon3:c.3480_3480delinsAT
4	106,158,595	NA	T	A	54	22	NA	0	*TET2*	NM_017628:exon3:c.T3496A:p.X1166K
9	5,073,770	rs386626619	G	T	101	20	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	115	14	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	117	18	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	125	11	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	126	14	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	126	21	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	175	16	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	31	59	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	45	56	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	47	73	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	49	57	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	63	53	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	64	17	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	66	23	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	69	15	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	70	9	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	73	10	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	79	9	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	81	7	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	81	9	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	84	13	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	87	19	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	88	28	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
9	5,073,770	rs386626619	G	T	88	42	12600	30,687	*JAK2*	NM_004972:exon14:c.G1849T:p.V617F
11	108,236,087	NA	G	A	81	7	1139600 21626	8	*ATM*	NM_000051:exon63:c.G9023A:p.R3008H
11	119,148,891	rs267606706	T	C	30	8	34052	24	*CBL*	NM_005188:exon8:c.T1111C:p.Y371H
11	119,149,251	rs267606708	G	A	109	18	34077	11	*CBL*	NM_005188:exon9:c.G1259A:p.R420Q
11	119,149,251	rs267606708	G	A	125	13	34077	11	*CBL*	NM_005188:exon9:c.G1259A:p.R420Q
15	90,631,935	NA	G	A	81	11	41877	10	*IDH2*	NM_002168:exon4:c.C418T:p.R140W
17	7,577,538	rs11540652	C	T	79	25	10662 1640830	71	*TP53*	NM_000546:exon7:c.G743A:p.R248Q

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
							3356964 99020 99021 99602			
17	7,577,538	rs11540652	C	T	83	15	10662 1640830 3356964 99020 99021 99602	71	*TP53*	NM_000546:exon7:c.G743A:p.R248Q
17	7,577,568	NA	C	T	63	29	11059 1649400 179811 179812 179813 3388191	8	*TP53*	NM_000546:exon7:c.G713A:p.C238Y
17	7,578,190	NA	T	C	26	17	10758 1644277 3355993 99718 99719 99720	23	*TP53*	NM_000546:exon6:c.A659G:p.Y220C
17	40,474,482	NA	T	A	188	18	1155743	45	*STAT3*	NM_003150:exon21:c.A1919T:p.Y640F
17	58,678,121	NA	G	GC	11	5	NA	0	*PPM1D*	NM_003620:exon1:c.346_346delinsGC
17	58,725,309	NA	GAC	G	37	41	NA	0	*PPM1D*	NM_003620:exon4:c.883_885G
17	58,734,163	NA	T	A	68	31	NA	0	*PPM1D*	NM_003620:exon5:c.T1221A:p.C407X
17	58,740,374	NA	TG	T	106	22	NA	0	*PPM1D*	NM_003620:exon6:c.1279_1280T
17	58,740,467	NA	C	T	42	37	NA	0	*PPM1D*	NM_003620:exon6:c.C1372T:p.R458X
17	58,740,467	NA	C	T	73	55	NA	0	*PPM1D*	NM_003620:exon6:c.C1372T:p.R458X
17	58,740,507	NA	CA	C	98	31	NA	0	*PPM1D*	NM_003620:exon6:c.1412_1413C
17	58,740,525	NA	AT	A	82	32	NA	0	*PPM1D*	NM_003620:exon6:c.1430_1431A
17	58,740,532	NA	T	TA	40	66	NA	0	*PPM1D*	NM_003620:exon6:c.1437_1437delinsTA
17	58,740,543	NA	C	CT	97	31	NA	0	*PPM1D*	NM_003620:exon6:c.1448_1448delinsCT
17	58,740,560	NA	TC	T	79	18	NA	0	*PPM1D*	NM_003620:exon6:c.1465_1466T
17	58,740,623	NA	C	CA	71	21	NA	0	*PPM1D*	NM_003620:exon6:c.1528_1528delinsCA
17	58,740,668	NA	G	T	62	19	982224	0	*PPM1D*	NM_003620:exon6:c.G1573T:p.E525X
17	58,740,713	NA	G	T	47	12	NA	0	*PPM1D*	NM_003620:exon6:c.G1618T:p.E540X
17	58,740,809	NA	C	T	60	10	NA	0	*PPM1D*	NM_003620:exon6:c.C1714T:p.R572X
17	74,732,935	NA	CGGCGGCTGTGGTGTGAGTCCGGGG	C	30	6	1318446 146289	23	*SRSF2*	NM_003016:exon1:c.284_308G
17	74,732,935	NA	CGGCGGCTGTGGTGTGAGTCCGGGG	C	86	9	1318446 146289	23	*SRSF2*	NM_003016:exon1:c.284_308G
17	74,732,959	NA	G	C	41	22	211661	30	*SRSF2*	NM_003016:exon1:c.C284G:p.P95R
17	74,732,959	NA	G	C	48	19	211661	30	*SRSF2*	NM_003016:exon1:c.C284G:p.P95R
17	74,732,959	NA	G	C	50	19	211661	30	*SRSF2*	NM_003016:exon1:c.C284G:p.P95R
17	74,732,959	NA	G	T	34	15	211029 211504 211505	84	*SRSF2*	NM_003016:exon1:c.C284A:p.P95H
17	74,732,959	NA	G	T	37	10	211029 211504 211505	84	*SRSF2*	NM_003016:exon1:c.C284A:p.P95H
20	31,019,423	NA	CA	C	35	30	NA	0	*ASXL1*	NM_015338:exon9:c.920_921C
20	31,021,158	NA	T	A	52	14	NA	0	*ASXL1*	NM_015338:exon11:c.T1157A:p.L386X
20	31,021,295	NA	C	T	71	21	NA	0	*ASXL1*	NM_015338:exon11:c.C1294T:p.Q432X
20	31,021,542	NA	CTG	C	194	33	NA	0	*ASXL1*	NM_015338:exon11:c.1541_1543C
20	31,021,565	NA	C	T	160	104	NA	0	*ASXL1*	NM_015338:exon11:c.C1564T:p.Q522X
20	31,021,622	NA	C	CGGCT	170	25	NA	0	*ASXL1*	NM_015338:exon11:c.1621_1621delinsCGGCT

Chromosome	Position (GRCh37)	dbSNP 138 ID	Reference Allele	Alternate Allele	Reference Count	Alternate Count	COSMIC ID	COSMIC Count	Gene	Annotation
20	31,022,286	NA	T	TA	74	15	36166	9	*ASXL1*	NM_015338:exon12:c.1771_1771delinsTA
20	31,022,402	NA	TCACCACTG CCATAGAGA GGCGGC	T	12	6	36165 41597 51200	61	*ASXL1*	NM_015338:exon12:c.1887_1910T
20	31,022,402	NA	TCACCACTG CCATAGAGA GGCGGC	T	13	7	36165 41597 51200	61	*ASXL1*	NM_015338:exon12:c.1887_1910T
20	31,022,402	NA	TCACCACTG CCATAGAGA GGCGGC	T	16	8	36165 41597 51200	61	*ASXL1*	NM_015338:exon12:c.1887_1910T
20	31,022,402	NA	TCACCACTG CCATAGAGA GGCGGC	T	29	3	36165 41597 51200	61	*ASXL1*	NM_015338:exon12:c.1887_1910T
20	31,022,402	NA	TCACCACTG CCATAGAGA GGCGGC	T	29	3	36165 41597 51200	61	*ASXL1*	NM_015338:exon12:c.1887_1910T
20	31,022,402	NA	TCACCACTG CCATAGAGA GGCGGC	T	30	5	36165 41597 51200	61	*ASXL1*	NM_015338:exon12:c.1887_1910T
20	31,022,402	NA	TCACCACTG CCATAGAGA GGCGGC	T	39	8	36165 41597 51200	61	*ASXL1*	NM_015338:exon12:c.1887_1910T
20	31,022,414	NA	TAG	T	14	6	NA	0	*ASXL1*	NM_015338:exon12:c.1899_1901T
20	31,022,485	NA	A	AG	7	4	NA	0	*ASXL1*	NM_015338:exon12:c.1970_1970delinsAG
20	31,022,572	NA	AGT	A	35	9	146261	2	*ASXL1*	NM_015338:exon12:c.2057_2059A
20	31,022,592	rs373221034	C	T	30	5	51388	11	*ASXL1*	NM_015338:exon12:c.C2077T:p.R693X
20	31,022,592	rs373221034	C	T	38	5	51388	11	*ASXL1*	NM_015338:exon12:c.C2077T:p.R693X
20	31,022,624	NA	TG	T	43	11	266052	0	*ASXL1*	NM_015338:exon12:c.2109_2110T
20	31,022,624	NA	T	TC	60	14	1155825	1	*ASXL1*	NM_015338:exon12:c.2109_2109delinsTC
20	31,022,688	NA	A	T	24	8	NA	0	*ASXL1*	NM_015338:exon12:c.A2173T:p.R725X
20	31,022,708	NA	AC	A	30	10	NA	0	*ASXL1*	NM_015338:exon12:c.2193_2194A
20	31,022,898	NA	TC	T	39	11	1716903 34212	4	*ASXL1*	NM_015338:exon12:c.2383_2384T
20	31,022,922	NA	C	T	84	19	96380	1	*ASXL1*	NM_015338:exon12:c.C2407T:p.Q803X
20	31,022,981	NA	AT	A	96	71	NA	0	*ASXL1*	NM_015338:exon12:c.2466_2467A
20	31,022,991	NA	G	T	117	18	NA	0	*ASXL1*	NM_015338:exon12:c.G2476T:p.G826X
20	31,023,045	NA	A	AC	247	47	1411087 41712	1	*ASXL1*	NM_015338:exon12:c.2530_2530delinsAC
20	31,023,083	NA	C	A	306	65	NA	0	*ASXL1*	NM_015338:exon12:c.C2568A:p.C856X
20	31,023,209	NA	G	A	50	13	NA	0	*ASXL1*	NM_015338:exon12:c.G2694A:p.W898X
20	31,023,408	NA	C	T	52	14	267971	3	*ASXL1*	NM_015338:exon12:c.C2893T:p.R965X
20	31,023,473	NA	C	CGT	92	20	NA	0	*ASXL1*	NM_015338:exon12:c.2958_2958delinsCGT
20	31,023,717	NA	C	T	92	26	41715	4	*ASXL1*	NM_015338:exon12:c.C3202T:p.R1068X
20	31,024,273	NA	G	GC	40	38	NA	0	*ASXL1*	NM_015338:exon12:c.3758_3758delinsGC
20	31,025,057	NA	CAT	C	60	49	NA	0	*ASXL1*	NM_015338:exon12:c.4542_4544C
21	44,524,456	rs371769427	G	A	26	5	1142948 166866	33	*U2AF1*	NM_006758:exon2:c.C101T:p.S34F

Table S4

Cysteine mutations in the *DNMT3A* gene. *DNMT3A* mutations leading to the formation of new cysteine residues and predicted de novo disulfide bond formation.

Mutation	Number of subjects	Disulfide bonds	Disulfide Bond Score*
G543C	1	524-543	0.99676
S714C	3	541-714	0.99651
F732C	1	497-732	0.97115
Y735C	4	520-735	0.30687
R736C	1	520-736	0.99095
R749C	2	749-818	0.99843
F751C	1	524-751	0.99811
W753C	1	554-753	0.72528
R882C	6	494-882	0.8412
L889C	1	818-889	0.99797

* http://clavius.bc.edu/~clotelab/DiANNA/

Note: Catalytic ADD-Domain amino acids 472-610

Table S5

Counts for subjects with one putative somatic mutation and no candidate drivers (one mut.), subjects with exactly two putative somatic mutations and no candidate drivers (two muts.), subjects with clonal hematopoiesis with unknown drivers (CH-UD), subjects with clonal hematopoiesis with candidate drivers (CH-CD), and subjects with clonal hematopoiesis with candidate or unknown drivers (CH). Subjects were counted across all individuals for whom both age at sampling information and sequencing data of sufficient quality for detection of putative somatic mutations were available, with the exception of subject with CH-CD for whom only age at sampling information was required.

Age	one mut.	two muts.	CH-UD	CH-CD	CH
19-30	18/174	1/174	0/174	1/196	1/174
31-35	36/349	5/349	2/349	2/371	3/349
36-40	48/661	13/661	1/661	5/708	5/661
41-45	93/1081	15/1081	5/1081	6/1154	9/1081
46-50	120/1303	12/1303	5/1303	18/1378	22/1303
51-55	148/1597	28/1597	10/1597	26/1695	32/1597
56-60	190/1725	41/1725	19/1725	41/1815	58/1725
61-65	187/1608	40/1608	35/1608	56/1659	88/1608
66-70	141/1105	36/1105	32/1105	44/1140	76/1105
71-75	77/600	29/600	29/600	48/619	75/600
76-80	57/355	15/355	32/355	25/356	58/355
81-93	13/73	5/73	5/73	7/73	12/73

Table S6

Subjects with clonal hematopoiesis and a diagnosis of hematologic malignancy after DNA sampling. There were 37 subjects diagnosed with hematologic malignancies after DNA sampling. Of these, 15 had showed clonal hematopoiesis in their initial DNA sample. Diagnoses of hematologic malignancies in these subjects followed DNA sampling by an average of 17 months (range: 2–36 months). Subjects with additional sequence generated to identify the malignancy are highlighted in bold.

Subject			Mutations		First diagnosis	
Sex	Age	Died	Candidate drivers	Passengers	Months after	Type
Male	62	Yes	NA	3	32	Unspecified B-cell lymphoma, unspecified site
Male	64	No	NA	3	7	Multiple myeloma
Male	70	Yes	*SF3B1* p.K700E	3	20	Chronic lymphocytic leukemia of B-cell type
Female	63	No	NA	3	11	Chronic lymphocytic leukemia of B-cell type
Male	63	No	NA	10	9	Chronic lymphocytic leukemia of B-cell type
Female	**72**	**Yes**	***TP53* p.R248Q**	**3**	**34**	**Acute myeloblastic leukemia**[3]
Male	73	Yes	*SRSF2* p.P95H	6	21	Acute myeloblastic leukemia
Female	71	No	*SRSF2* p.P95H	1	9	Chronic myelomonocytic leukemia
Male	**64**	**No**	**NA**	**3**	**2**	**Acute leukemia of unspecified cell type**[2]
Female	73	Yes	*DNMT3A* p.V372A	0	36	Chronic leukemia of unspecified cell type
Female	61	Yes	*DNMT3A* p.P904L	1	11	Other myelodysplastic syndromes
Male	**85**	**Yes**	***SRSF2* p.P95H**	**13**	**2**	**Other myelodysplastic syndromes**[1]
Male	69	No	*JAK2* p.V617F	2	35	Chronic myeloproliferative disease
Female	76	No	*JAK2* p.V617F	4	13	Chronic myeloproliferative disease
Male	57	No	*DNMT3A* p.H613D	0	14	Monoclonal gammopathy

[1] Subject #1

[2] Subject #2 (later progressed to acute myeloblastic leukemia)

[3] Subject #3

Table S7

Subjects with clonal hematopoiesis and a diagnosis of hematologic malignancy before DNA sampling. There were 55 subjects with a previous diagnosis of hematologic malignancy up to 12 years before DNA sampling. Of these, 14 showed clonal hematopoiesis. Previous history of hematologic malignancy was a strong risk factor for clonal hematopoiesis (OR=6.0; 95% CI 3.1 to 12; P<0.001, adjusting for age and sex using a linear regression model).

Subject			Mutations		First diagnosis	
Sex	Age	Died	Candidate drivers	Passengers	Months before	Type
Female	64	No	NA	6	95	Hodgkin lymphoma, unspecified
Female	72	Yes	NA	18	148	Hodgkin lymphoma, unspecified
Female	72	No	*DNMT3A* p.R556G	2	17	Follicular lymphoma, unspecified
Male	63	No	*DNMT3A* p.R597W	0	12	Diffuse large B-cell lymphoma
Male	76	Yes	NA	3	52	Other non-follicular lymphoma, unspecified site
Male	61	No	*DNMT3A* p.E907K *PPM1D* frameshift	0	13	Other specified types of non-Hodgkin lymphoma
Female	61	No	*DNMT3A* p.G543C	2	145	Acute leukemia of unspecified cell type
Male	57	No	NA	3	1	Polycythemia vera
Male	51	No	*JAK2* p.V617F	3	49	Polycythemia vera
Male	70	No	*JAK2* p.V617F	1	46	Polycythemia vera
Male	61	No	*JAK2* p.V617F	1	25	Polycythemia vera
Male	77	Yes	*CBL* p.Y371H *U2AF1* p.S34F	9	46	Other myelodysplastic syndromes
Male	57	No	*JAK2* p.V617F	5	4	Chronic myeloproliferative disease
Female	56	No	*JAK2* p.V617F	0	20	Essential (hemorrhagic) thrombocythemia

Table S8

Subjects with clonal hematopoiesis at DNA sampling who died during follow-up. Subjects with additional sequence generated to identify the malignancy are highlighted in bold.

Subject		Mutations		Death	
Sex	Age	Candidate Drivers	Passengers	Months after	Cause
Male	73	NA	3	7	Malignant neoplasm of sigmoid colon
Male	67	*DNMT3A* p.Y908C	0	65	Malignant neoplasm of prostate
Male	74	*ASXL1* p.Q803X	1	30	Malignant neoplasm of prostate
Male	76	NA	3	17	Unspecified B-cell lymphoma
Female	72	NA	18	3	Unspecified Non-Hodgkin lymphoma
Female	61	*DNMT3A* p.P904L	1	18	Acute myeloblastic leukaemia [AML]
Female	**72**	***TP53* p.R248Q**	**3**	**36**	**Acute myeloblastic leukaemia [AML]**
Male	73	*SRSF2* p.P95R	6	26	Acute myeloblastic leukaemia [AML]
Male	**85**	***SRSF2* p.P95H**	**13**	**16**	**Unspecified leukemia**
Male	77	*CBL* p.Y371H *U2AF1* p.S34F	9	16	Myelodysplastic syndrome, unspecified
Male	78	NA	3	19	Anemia, unspecified
Male	63	*DNMT3A* frameshift	0	6	Haemophagocytic syndrome, infection-associated
Male	68	NA	4	14	Diabetes mellitus type 2 with renal complications
Male	76	NA	5	6	Unspecified diabetes mellitus with multiple complications
Male	59	*ASXL1* frameshift	0	6	Unspecified diabetes mellitus without complications
Male	72	*PPM1D* p.E540X	5	4	Parkinson disease
Male	66	NA	3	5	Anoxic brain damage, not elsewhere classified
Female	64	*JAK2* p.V617F	4	45	Acute myocardial infarction, unspecified
Female	82	NA	5	37	Acute myocardial infarction, unspecified
Male	59	*DNMT3A* p.E599D	0	30	Acute myocardial infarction, unspecified
Female	74	NA	3	9	Atherosclerotic heart disease
Female	64	*DNMT3A* p.F751C	2	40	Pulmonary heart disease, unspecified
Male	73	*SF3B1* p.K666T *TET2* frameshift	8	12	Acute and subacute infective endocarditis
Male	77	NA	5	10	Endocarditis, valve unspecified
Male	77	NA	7	27	Heart failure, unspecified
Female	80	NA	4	10	Heart failure, unspecified
Female	65	*PPM1D* p.R458X	0	10	Cardiomegaly
Female	75	*TET2* frameshift	2	19	Subarachnoid haemorrhage, unspecified
Female	88	*ASXL1* p.R965X	3	7	Intracerebral haemorrhage, unspecified
Male	67	NA	4	34	Stroke, not specified as haemorrhage or infarction
Female	64	*DNMT3A* p.C520X	3	24	Other specified cerebrovascular diseases
Male	81	*DNMT3A* p.L344P	0	42	Sequelae of other and unspecified cerebrovascular diseases
Female	70	*DNMT3A* p.R882H	0	48	Generalized and unspecified atherosclerosis
Male	54	NA	3	39	Generalized and unspecified atherosclerosis
Male	66	*DNMT3A* p.I681M	4	32	Generalized and unspecified atherosclerosis
Female	75	*DNMT3A* p.Q816X	0	7	Unspecified chronic bronchitis
Male	68	NA	3	11	Chronic obstructive pulmonary disease, unspecified
Female	62	ASXL1 frameshift	2	57	Chronic obstructive pulmonary disease, unspecified
Male	74	NA	5	39	Chronic obstructive pulmonary disease, unspecified
Female	65	*DNMT3A* p.E733G	7	26	Chronic obstructive pulmonary disease, unspecified

Subject		Mutations		Death	
Sex	Age	Candidate Drivers	Passengers	Months after	Cause
		JAK2 p.V617F			
Male	57	NA	3	34	Gastro-oesophageal reflux disease with oesophagitis
Male	51	*DNMT3A* p.M761V	6	28	Other ill-defined and unspecified causes of mortality
Female	65	*TET2* frameshift	3	15	Other ill-defined and unspecified causes of mortality
Male	77	NA	4	26	Unspecified drowning and submersion
Female	74	*DNMT3A* p.P307R	3	44	Unknown
Male	64	*SF3B1* p.K666N	7	48	Unknown
Male	70	NA	3	52	Unknown
Male	62	NA	3	43	Unknown
Male	70	*SF3B1* p.K700E	3	43	Unknown
Male	74	*ASXL1* frameshift	4	41	Unknown
Female	73	*DNMT3A* p.V372A	0	42	Unknown
Male	67	*JAK2* p.V617F	4	44	Unknown
Male	72	*IDH2* p.R140W SRSF2 frameshift	2	37	Unknown
Female	75	*DNMT3A* p.Y735C	0	27	Unknown

Table S9

Somatic mutations for Subject #1. List of putative somatic mutations and candidate driver somatic mutations from whole-exome sequencing (WES) data and high coverage whole-genome sequencing (WGS) data of blood. Candidate driver somatic mutations are highlighted in bold.

Subject #1 (diagnosed with myeloid malignancy 2 months after DNA sampling)

Chromosome	Position (GRCh37)	dbSNP 138 or COSMIC ID	Reference Allele	Alternate Allele	Reference Count (WES blood)	Alternate Count (WES blood)	Reference Count (WGS blood)	Alternate Count (WGS blood)	Gene	Annotation
1	197,070,852	NA	A	G	82	23	115	19	*ASPM*	NM_018136:exon18:c.T7529C:p.I2510T
2	242,178,077	NA	T	G	196	79	109	33	*HDLBP*	NM_005336:exon20:c.A2736C:p.R912S
3	38,519,942	NA	G	A	65	18	107	35	*ACVR2B*	NM_001106:exon5:c.G599A:p.R200H
3	46,306,703	NA	T	A	52	8	126	21	*CCR3*	NM_001837:exon3:c.T54A:p.D18E
3	52,437,754	rs150524807	G	A	52	8	124	31	*BAP1*	NM_004656:exon13:c.C1407T:p.S469S
4	**106,162,527**	**NA**	**T**	**TTA**	**0**	**0**	**111**	**17**	***TET2***	**NM_001127208:exon4:c.3441_3441delinsTTA**
4	**106,164,929**	**NA**	**A**	**G**	**0**	**0**	**126**	**22**	***TET2***	**NM_001127208:exon6:c.A3797G:p.N1266S**
4	158,284,236	NA	C	T	79	21	107	22	*GRIA2*	NA
5	54,404,054	NA	G	A	74	10	108	23	*GZMA*	NM_006144:exon4:c.G459A:p.W153X
6	50,696,983	COSM3354285	C	T	160	44	105	28	*TFAP2D*	NM_172238:exon5:c.C841T:p.R281W
11	67,265,009	NA	C	T	198	25	146	18	*PITPNM1*	NM_004910:exon13:c.G1924A:p.E642K
13	23,909,533	rs9552930	T	C	75	17	107	38	*SACS*	NM_014363:exon10:c.A8482G:p.S2828G
14	92,472,207	NA	G	C	154	30	118	16	*TRIP11*	NM_004239:exon11:c.C2113G:p.L705V
15	43,668,387	NA	A	T	110	32	139	31	*TUBGCP4*	NM_014444:exon2:c.A170T:p.E57V
17	**74,732,959**	**COSM211029** **COSM211504** **COSM211505**	**G**	**T**	**37**	**10**	**100**	**20**	***SRSF2***	**NM_003016:exon1:c.C284A:p.P95H**
20	1,107,965	NA	A	G	196	29	103	20	*PSMF1*	NA
20	**31,022,441**	**COSM34210** **COSM1411076** **COSM1658769**	**A**	**AG**	**10**	**3**	**89**	**31**	***ASXL1***	**NM_015338:exon12:c.1926_1926delinsAG**
21	**36,259,198**	**COSM24719** **COSM24728**	**AG**	**A**	**57**	**3**	**127**	**17**	***RUNX1***	**NM_001754:exon4:c.292_293T**
X	**123,191,828**	**NA**	**G**	**A**	**28**	**2**	**33**	**12**	***STAG2***	**NM_001042750:exon15:c.1416+1G>A**

Table S10

Somatic mutations for Subject #2. List of putative somatic mutations and candidate driver somatic mutations from whole-exome sequencing (WES) data of blood, high coverage whole-genome sequencing (WGS) data of blood, and whole-exome sequencing data for bone marrow biopsy at the time of first diagnosis. Candidate driver somatic mutations are highlighted in bold.

Subject #2 (diagnosed with AML 2 months after DNA sampling)

Chromosome	Position (GRCh37)	dbSNP 138 or COSMIC ID	Reference Allele	Alternate Allele	Reference Count (WES blood)	Alternate Count (WES blood)	Reference Count (WGS blood)	Alternate Count (WGS blood)	Reference Count (WES bone marrow)	Alternate Count (WES bone marrow)	Gene	Annotation
11	123,811,251	NA	G	A	91	20	150	11	36	14	*OR4D5*	NM_001001965:exon1:c.G928A:p.G310S
19	10,090,052	NA	G	A	182	32	149	14	140	38	*COL5A3*	NM_015719:exon38:c.C2754T:p.V918V
19	**33,792,380**	**COSM27466**	**A**	**ACCTTCTGCTGCGTCTCCACGTTGCGCTGCTTGG**	**42**	**0**[*]	**55**	**13**[*]	**85**	**7**[*]	***CEBPA***	**NM_004364:exon1:c.941_941delinsCCAAGCAGCGCAACGTGGAGACGCAGCAGAAGGT**
19	**33,793,111**	**COSM18539** **COSM29127** **COSM29220**	**CG**	**C**	**0**	**0**	**92**	**7**[^]	**26**	**3**[^]	***CEBPA***	**NM_004364:exon1:c.210_211G**
20	43,129,883	NA	C	T	109	18	136	16	110	22	*SERINC3*	NM_006811:exon9:c.G1114A:p.V372I

[*] due to the size of this insertion, alternate allele count is dependent on sequencing reads length, 76 for WES blood, 151 for WGS blood, and 101 for WES bone marrow

[^] this mutation was not automatically genotyped by the Haplotype Caller from the Genome Analysis Toolkit due to low allelic count

Table S11

Somatic mutations for Subject #3. List of putative somatic mutations and candidate driver somatic mutations from whole-exome sequencing (WES) data of blood and whole-exome sequencing data for bone marrow biopsy at the time of first diagnosis. Candidate drivers are highlighted in bold.

Subject #3 (diagnosed with AML 34 months after DNA sampling)

Chromosome	Position (GRCh37)	dbSNP 138 or COSMIC ID	Reference Allele	Alternate Allele	Reference Count (WES blood)	Alternate Count (WES blood)	Reference Count (WES bone marrow)	Alternate Count (WES bone marrow)	Gene	Annotation
3	126,178,475	NA	C	T	277	87	100	95	*ZXDC*	NA
4	154,523,476	NA	C	T	181	37	167	2	*KIAA0922*	NM_015196:exon22:c.C2436T:p.R812R
17	**7,577,538**	**rs11540652 COSM10662 COSM99020 COSM99021 COSM99602 COSM1640830 COSM3356964**	**C**	**T**	**79**	**25**	**16**	**101**	***TP53***	**NM_000546:exon7:c.G743A:p.R248Q**
17	72,297,215	NA	G	C	156	27	62	4	*DNAI2*	NM_023036:exon8:c.G895C:p.E299Q
5		**del(5q)**	**NA**	**NA**	**~8%**		**~86%**		**NA**	**NA**
17		**del(17)**	**NA**	**NA**	**~3%**		**~86%**		**NA**	**NA**
12,13,16,19		**complex karyotype**	**NA**	**NA**	**~0%**		**~86%**		**NA**	**NA**

References

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
2. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297 (2010).
3. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–950 (2011).
4. Abdel-Wahab, O., Kilpivaara, O., Patel, J., Busque, L. & Levine, R. L. The most commonly reported variant in ASXL1 (c.1934dupG;p.Gly646TrpfsX12) is not a somatic alteration. *Leukemia* **24,** 1656–1657 (2010).
5. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinforma. Oxf. Engl.* (2014). doi:10.1093/bioinformatics/btu356
6. Genovese, G., Handsaker, R. E., Li, H., Kenny, E. E. & McCarroll, S. A. Mapping the Human Reference Genome's Missing Sequence by Three-Way Admixture in Latino Genomes. *Am. J. Hum. Genet.* **93,** 411–421 (2013).
7. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).
8. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
9. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
10. Langemeijer, S. M. C. *et al.* Acquired mutations in TET2 are common in myelodysplastic syndromes. *Nat. Genet.* **41,** 838–842 (2009).
11. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Prepr. ArXiv13033997* (2013).

12. Hindson, B. J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83,** 8604–8610 (2011).

13. Kralovics, R. *et al.* A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N. Engl. J. Med.* **352,** 1779–1790 (2005).

14. Marková, J. *et al.* Prognostic impact of DNMT3A mutations in patients with intermediate cytogenetic risk profile acute myeloid leukemia. *Eur. J. Haematol.* **88,** 128–135 (2012).

15. Roller, A. *et al.* Landmark analysis of DNMT3A mutations in hematological malignancies. *Leukemia* **27,** 1573–1578 (2013).

16. Gaidzik, V. I. *et al.* Clinical impact of DNMT3A mutations in younger adult patients with acute myeloid leukemia: results of the AML Study Group (AMLSG). *Blood* **121,** 4769–4777 (2013).

17. Russler-Germain, D. A. *et al.* The R882H DNMT3A Mutation Associated with AML Dominantly Inhibits Wild-Type DNMT3A by Blocking Its Ability to Form Active Tetramers. *Cancer Cell* **25,** 442–454 (2014).

18. Jurkowska, R. Z. *et al.* Oligomerization and binding of the Dnmt3a DNA methyltransferase to parallel DNA molecules: heterochromatic localization and role of Dnmt3L. *J. Biol. Chem.* **286,** 24200–24207 (2011).

19. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

20. Ferrè, F. & Clote, P. DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res.* **33,** W230–232 (2005).

21. Zhang, Y. *et al.* Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Res.* **38,** 4246–4253 (2010).

22. Kelley, L. A. & Sternberg, M. J. E. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4,** 363–371 (2009).

23. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46,** 624–628 (2014).

24. Stone, J. F. & Sandberg, A. A. Sex chromosome aneuploidy and aging. *Mutat. Res.* **338,** 107–

113 (1995).

25. Du, Y., Fryzek, J., Sekeres, M. A. & Taioli, E. Smoking and alcohol intake as risk factors for myelodysplastic syndromes (MDS). *Leuk. Res.* **34,** 1–5 (2010).

26. Strom, S. S., Oum, R., Elhor Gbito, K. Y., Garcia-Manero, G. & Yamamura, Y. De novo acute myeloid leukemia risk factors: a Texas case-control study. *Cancer* **118,** 4589–4596 (2012).

27. Fircanis, S., Merriam, P., Khan, N. & Castillo, J. J. The relation between cigarette smoking and risk of acute myeloid leukemia: An updated meta-analysis of epidemiological studies. *Am. J. Hematol.* **89,** E125–132 (2014).

28. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–251 (2014).

29. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481,** 506–510 (2012).

30. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).

31. Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N. Engl. J. Med.* **366,** 1090–1098 (2012).

32. Walter, M. J. *et al.* Clonal diversity of recurrently mutated genes in myelodysplastic syndromes. *Leukemia* **27,** 1275–1282 (2013).

33. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).

34. Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res.* **24,** 733–742 (2014).

35. Chen, T.-C. *et al.* Dynamics of ASXL1 mutation and other associated genetic alterations during disease progression in patients with primary myelodysplastic syndrome. *Blood Cancer J.* **4,** e177 (2014).

36. Lin, L.-I. *et al.* A novel fluorescence-based multiplex PCR assay for rapid simultaneous detection of CEBPA mutations and NPM mutations in patients with acute myeloid leukemias.

Leukemia **20,** 1899–1903 (2006).

37. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

38. Barjesteh van Waalwijk van Doorn-Khosrovani, S. *et al.* Biallelic mutations in the CEBPA gene and low CEBPA expression levels as prognostic markers in intermediate-risk AML. *Hematol. J. Off. J. Eur. Haematol. Assoc. EHA* **4,** 31–40 (2003).

39. Soenen-Cornu, V., Preudhomme, C., Laï, J., Zandecki, M. & Fenaux, P. del(17p) in myeloid malignancies. *Atlas Genet. Cytogenet. Oncol. Haematol.* (2011). doi:10.4267/2042/37563

40. Kanehira, K., Ketterling, R. & Van, D. D. del(5q) in myeloid neoplasms. *Atlas Genet. Cytogenet. Oncol. Haematol.* (2011). doi:10.4267/2042/44718

41. Kulasekararaj, A. G. *et al.* TP53 mutations in myelodysplastic syndrome are strongly correlated with aberrations of chromosome 5, and correlate with adverse prognosis. *Br. J. Haematol.* **160,** 660–672 (2013).