Supplementary methods

Exome sequencing and data analysis

1. Sequencing

Detailed information about the sequenced samples is described in Supplemental Table 1 and Supplemental Figure 1. Genomic DNA from patients' bone marrow, peripheral blood, and skin tissue-derived fibroblast were submitted to NIH Intramural Sequencing Center (NISC) for exome sequencing. In general, gDNA was fragmented with Covaris Focused-ultrasonicator into small fragments around 300 bp size. Swift Biosciences Accel-NGS 2S Plus DNA Library Kit was used to perform end repair, A-tailing, adapter ligation, and a limited cycle of amplification, following the standard protocol. 500 ng of each of the eight libraries were pooled together, following exome capture procedurals with IDT xGen Exome Research Panel. NimbleGen SeqCap EZ Hybridization and Wash Kit (Roche) were used for a few archived samples. Prepared libraries were sequenced on Illumina Novaseq 6000 platform with a paired-end strategy aiming to achieve 200X average depth. In addition, several archive data (EMKFPD samples) were captured with Roche NimbleGen SeqCap EZ Hybridization and Wash Kit and sequenced for about 80X average depth.

2. Alignment and recalibration

Sequencing data were transferred to NIH high-performance computing system "Biowulf", and analyzed with an in-house pipeline. First of all, we used fastp[1] to do quality control of fastq files with default parameters. Low-quality reads and sequencing adaptor sequences were removed. Clean sequencing reads were mapped to hg38 reference with a streamed pipe of bwa[2] (v.0.7.17) alignment, sambamba[3] (v.0.8.0) sort bam files, and picard MarkDuplicates mark duplicates (v.2.17.11). Since most libraries were sequenced in multiple lanes, bam files of different lane's data were merged with sambamba. BQSR recalibration was further performed with GATK best practice[4] (v.4.2.0.0) to generate the analysis-ready bam file.

3. Genotype analysis

First of all, we used GATK (v.4.2.0.0) to do joint genotyping. In brief, per sample genotype were identified by HaplogypeCaller module with parameters "--emit-ref-confidence GVCF --max-alternate-alleles 3 --read-filter OverclippedReadFilter", and validated by ValidateVariants module with parameters "-gvcf --validation-type-to-exclude ALLELES".

All gvcf results of individual samples were used as input of GenomicsDBImport module, and followed by steps of joint genotyping. Including GenotypeGVCFs with parameters "-G StandardAnnotation -G AS_StandardAnnotation --allow-old-rms-mapping-quality-annotation-data --merge-input-intervals"; and VariantFiltration module with parameters "--filter-expression "ExcessHet>54.69" --filter-name ExcessHet". We further performed variant recalibration for SNPs and InDels.

For InDels, recalibrate model were generated by VariantRecalibrator model with parameters "--trust-all-polymorphic -tranche 100.0 -tranche 99.95 -tranche 99.9 -tranche 99.5 -tranche 99.0 -tranche 97.0 -tranche 96.0 -tranche 95.0 -tranche 94.0 -tranche 93.5 -tranche 93.0 -tranche 92.0 -tranche 91.0 -tranche 90.0 -an FS -an ReadPosRankSum -an MQRankSum -an QD -an SOR -an DP --use-allele-specific-annotations -mode INDEL --max-gaussians 4 --resource:mills,known=false,training=true,truth=true,prior=12 --resource:axiomPoly,known=false,training=true,truth=false,prior=10 --resource:dbsnp,known=true,training=false,truth=false,prior=2", and the model was applied to InDels by ApplyVQSR module with parameters "--use-allele-specific-annotations --truth-sensitivity-filter-level 95.0 --create-output-variant-index true -mode INDEL".

For SNPs, we first generated model by VariantRecalibrator module with parameters "--trust-all-polymorphic -tranche 100.0 -tranche 99.95 -tranche 99.9 -tranche 99.8 -tranche 99.6 -tranche 99.5 -tranche 99.4 -tranche 99.3 -tranche 99.0 -tranche 98.0 -tranche 97.0 -tranche 90.0 -an QD -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an SOR -an DP --use-allele-specific-annotations -mode SNP --output-model module.file.path --sample-every-Nth-variant 10 --max-gaussians 6 -resource:hapmap,known=false,training=true,truth=true,prior=15 -resource:omni,known=false,training=true,truth=true,prior=12 -resource:1000G,known=false,training=true,truth=false,prior=10 -resource:dbsnp,known=true,training=false,truth=false,prior=7". This model were feed as input for the second round VariantRecalibrator to generate VQSR model with parameters "--trust-all-polymorphic -tranche 100.0 -tranche 99.95 -tranche 99.9 -tranche 99.8 -tranche 99.6 -tranche 99.5 -tranche 99.4 -tranche 99.3 -tranche 99.0 -tranche 98.0 -tranche 97.0 -tranche 90.0 -an QD -an MQRankSum -an ReadPosRankSum -an FS -an MQ -an SOR -an DP --use-allele-specific-annotations -mode SNP --input-model module.file.path --max-gaussians 6 -resource:hapmap,known=false,training=true,truth=true,prior=15 -resource:omni,known=false,training=true,truth=true,prior=12 -resource:1000G,known=false,training=true,truth=false,prior=10 -resource:dbsnp,known=true,training=false,truth=false,prior=7". Finally, the VQSR model

were used in ApplyVQSR module to generate the final VQSR calibrated SNPs. Variants with PASS tag and QUAL>100 were kept.

To identify possible missing variants, especially low VAF variants, we also used freebayes[5] (v.1.3.5) and LoFreq[6] (v.2.1.5) to do genotyping as supplementary results. In Freebayes analysis, default parameters were used, following filtration criteria including (QUAL>20, DP>8, QUAL/AO>10, SAF>0, SAR>0, RPR>1, RPL>1). LoFreq analysis was also used with default parameters, mutations didn't show in GATK, and freebayes results with PASS tag were kept.

Genotype matrix were combined from three software, then annotated with SNPEff[7] and SNPSift. Multiple filtration procedurals were used to filter unreliable or polymorphism. First, mutations were compared with a region filter. Mutations located in exclude region files were removed, including DAC list, DUKE list, LCR list, Homopolymer list, and Mappable issue list. In addition, only if variants located in the exome capture region with 150bp flank on both ends, 1000G mappable region, and GIAB mappable regions, the variants were kept. Second, we did VCF normalization for all the variants passed region filters, to ensure the parsimony and left alignment of variants. Third, the population frequency of 1000G project, ExAC database, and gnomAD3 database were used to annotate all variants. Variants with 1000G VAF<1% and ExAC VAF<1%, and gnomAD3 popmax<1% were kept. In the last, all variants were manually reviewed with IGVtools to minimize false positive results.

For all variants that passed all filters, we re-evaluated depth, VAF, DP4 counts, based on high-quality sequencing reads (sequence_length >= 80, not secondary_alignment, and mapping_quality>50), and generated the finalized genotype matrix.

4. Somatic mutation calling

Somatic mutation identification is important to studying clonal hematopoiesis in FPDMM patients, and it's the key to discovering determinate factors of disease development. To improve the sensitivity of somatic mutation calling, we applied several software to identify somatic mutations independently and merged the mutation calling set together. To distinguish somatic mutations in bone marrow or peripheral blood cells, we collected skin biopsies from patients and expanded fibroblast, and further sequenced them as somatic control. However, since the fibroblast preparation procedure takes additional time, and it could fail, we have several patients who don't have skin fibroblast sequencing data ready for analysis. So, two somatic mutation analysis strategies were used in our work. For patients with fibroblast control data, we define the somatic mutations calling results as true somatic. For other patients without fibroblast control data, we have to use a single

sample somatic mutation calling strategy. Since we cannot confirm the existence in other tissues, we define the somatic mutation calling set as likely somatic.

True somatic mutation calling strategy: Base Recalibrated bam files of bone marrow/peripheral blood sample and fibroblast sample were used as input for Mutect2[4] (v.4.2.0.0), strelka[8] (v.2.9.0), LoFreq[6] (v.2.1.5), and MuSE[9] (v.1.0).

For Mutect2, we used 10 family controls without *RUNX1* variant to generate panel of normal (PoN) for analysis. Mutect2 module took bam files, reference, panel of normal, gnomAD germline resource to detect somatic mutations, and generate orientation bias model file. The orientation bias model was further processed by LearnReadOrientationModel, GetPileupSummaries, and CalculateContamination with default parameters. Finally, the processed bias model was used to filter the somatic mutation list. Mutations that didn't pass the model were removed from the analysis.

To analyze somatic mutation with Strelka, paired bam files were firstly used by Manta (v.1.5.0) to identify candidate small indels. After that, paired bam files, and candidate small indels reported by Manta were used by Strelka to report somatic SNVs and InDels. Default parameters were used with both Manta and Strelka.

We also used LoFreq, and MuSE2 to identify somatic SNVs and InDels with default parameters. Finally, we queried the genotype matrix to compare the genotype of BM/PB sample to the fibroblast sample, and generated the merged true somatic mutation list.

Likely somatic mutation calling strategy: Mutect2 single sample mode was used to predict somatic mutation from BM/PB sample itself. Previous described panel of normal (PoN) was used in this step to compare with. Similar to sample-control Mutect2 analysis, LearnReadOrientationModel, GetPileupSummaries, and CalculateContamination were done and filtered the somatic mutations. Since we sequenced multiple affected and unaffected individuals from the same family, if the mutation can be found in any other family members, we treat it as germline variant.

5. Somatic mutation filtering process

Similar to the genotyping process, mutations located in the region from DAC list, DUKE list, LCR list, Homopolymer list, and Mappable issue list were removed, we only kept mutations found in the exome capture region and flanking 150bp region in both upstream and downstream. In addition, to further remove mutations located in the low-complexity region, if the surrounding sequence met the criteria like: BBBBB, ABABABAB, ABCABCABC, ABCDABCD, mutations will be removed.

For the mutation evidence, the mutation locus must met "depth >=10, VAF>=0.03, and alt-allele-reads>=3" in somatic sample, and "depth >=8, VAF<0.01" in fibroblast control. To remove the strand-bias, the alt-allele-reads must include at least one supporting read from both chains. Because we would like to detect early events of somatic mutations in CHIP-related genes and leukemia driver genes, we called back all somatic mutations in these genes that failed the general filters, and reviewed the alignment carefully. If the somatic sample met "VAF>0.01 and alt-allele-reads>=3", and the alignment is clean, we kept these mutations in downstream analysis. However, to do the comparative analysis with the TOPMed study, we only included somatic mutations with VAF>0.05 in the provided 74 gene list from the TOPMed study.

6. In silico prediction of deleterious germline variants

To predict deleterious status of germline variants in our cohort, the following inclusion criteria were used to find out variants that potentially related to functional changes that could further impact the patients. Variants with any of these criteria will be included:

1) Germline variants that have been annotated as “Pathogenic/Likely pathogenic” with the CLINVAR database. For the variants related to Fig S9B, if ClinVar's preferred disease name (CLNDN) refers to Fanconi anemia, variants with uncertain significance or conflicting interpretations of pathogenicity will be counted.
2) Variants that are annotated to the following categories: “stop_gained/ frameshift_variant”.
3) Variants that are annotated to the following categories: “splice_donor_variant/ splice_acceptor_variant/splice_region_variant, meanwhile the SpliceAI annotation shows any of the DS_AG, DS_AL, DS_DG, or DS_DL values above 0.5.
4) Variants for which at least 4 out of 6 in-silico prediction tools support a deleterious outcome. These tools include CADD_phred>20; SIFT_pred=’D’; FATHMM_pred=’D’; Polyphen2_HDIV_pred=’D’ or ‘P’; MutationTaster_pred=’D’ or ‘A’; MetaSVM_pred=’D’.

We did the analysis on all germline variants to get the subset of variants that "predicted deleterious", then we put all the genes from this subset, performed enrichment analysis with Metascape[10]. Functional annotation categories from Gene Ontology, KEGG pathways, WikiPathways, and reactome were aggregated into four groups:

1. Hemostasis (R-HSA-109582, GO:0007599)

2. Coagulation and clotting (GO:0007596, GO:0050817, WP272, WP272, hsa04610, ko04610, GO:0030193, WP558, R-HSA-140877)
3. Platelet maturation and activation (R-HSA-76002, R-HSA-114608, GO:0030168, GO:0070527, GO:0010543, GO:0090330)
4. Myeloid cell differentiation (GO:0030099, GO:0002573, GO:0045637, GO:0045639, GO:0061515, GO:0002761, GO:0002763).

Genes related to RTK-RAS-PI3K pathway and Fanconi anemia were plotted and listed in the supplemental table 5.

7. Other analysis.

We used Maftools to do the mutation burden analysis[10]. And we used SigProfilerMatrixGenerator[11] to do the somatic mutation signature analysis. T-test was used to evaluate the significance in the genotype-phenotype analysis.

RNA sequencing and data analysis

1. Sequencing

RNA was extracted from PB or BM samples collected with PAXgene Bone Marrow/Blood DNA Kit (Qiagen) or from isolated PB/BM mononuclear cells. We took 0.5-1 ug of total bone marrow or peripheral blood RNA as input of RNA-seq. The RNA quality must meet the RIN>7.

RNA-seq libraries were prepared by NIH Intramural Sequencing Center (NISC) with Illumina TruSeq Stranded Total RNA with Ribo-Zero Globin strategy, following by PE151 sequencing on Novaseq 6000 platform. In addition, libraries were sequenced on the Illumina Novaseq platform with PE151 chemistry.

2. Analysis

Sequencing data were analyzed with an in-house pipeline under the snakemake pipeline framework. First of all, low-quality reads and sequencing adaptor sequences were removed by fastp with default parameters. STAR[12] (v.2.7.6a) was used to align all clean sequencing reads to human reference genome hg38, with two pass alignment strategy. Specific parameters were used in pass 1 (--outFilterMultimapScoreRange 1 --outFilterMultimapNmax 20 --outFilterMismatchNmax 10 --alignIntronMax 500000 --alignMatesGapMax 1000000 --sjdbScore 2 --alignSJDBoverhangMin 1 --

outFilterMatchNminOverLread 0.33 --outFilterScoreMinOverLread 0.33), and pass 2 (--outFilterMultimapScoreRange 1 --outFilterMultimapNmax 20 --outFilterMismatchNmax 10 --alignIntronMax 500000 --alignMatesGapMax 1000000 --sjdbScore 2 --alignSJDBoverhangMin 1). Aligned bam files were then sorted and indexed with sambamba[3] (v.0.8.0). Splicing junction forms and counts were collected from the output of STAR.

We applied STARfusion[12] (v1.8.0) and Arriba[13] (v1.2.0) in the fusion gene analysis to maximize the detection of fusion events, followed the standard protocols.

SNP array tests and data analysis

For each sample, 300 ng high-quality genomic DNA was processed and loaded with OmniExpressExome-8 v1.6 bead chip. The data was processed with GenomeStudio, and CNV analysis was performed with two analysis plug-ins: cnvPartition and PennCNV[14]. We also illustrated the copy number Log R ratio (LRR) and B allele frequency (BAF) with an in-house script. All predicted CNV events were manually reviewed to eliminate false positive calls.

Other

We developed standardized exome genotyping, somatic mutation calling, and RNA-seq analysis pipelines with the snakemake framework[15]. All customized figures in this manuscript were generated with matplotlib.

References

1 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

2 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

3 Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034, doi:10.1093/bioinformatics/btv098 (2015).

4 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
5 Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* (2012).
6 Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* **40**, 11189-11201, doi:10.1093/nar/gks918 (2012).
7 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92, doi:10.4161/fly.19695.
8 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594, doi:10.1038/s41592-018-0051-x (2018).
9 Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol* **17**, 178, doi:10.1186/s13059-016-1029-6 (2016).
10 Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res* **28**, 1747-1756, doi:10.1101/gr.239244.118 (2018).
11 Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685, doi:10.1186/s12864-019-6041-2 (2019).
12 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
13 Uhrig, S. *et al.* Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res* **31**, 448-460, doi:10.1101/gr.257246.119 (2021).
14 Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-1674, doi:10.1101/gr.6861907 (2007).
15 Molder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, 33, doi:10.12688/f1000research.29032.2 (2021).

Figure S1

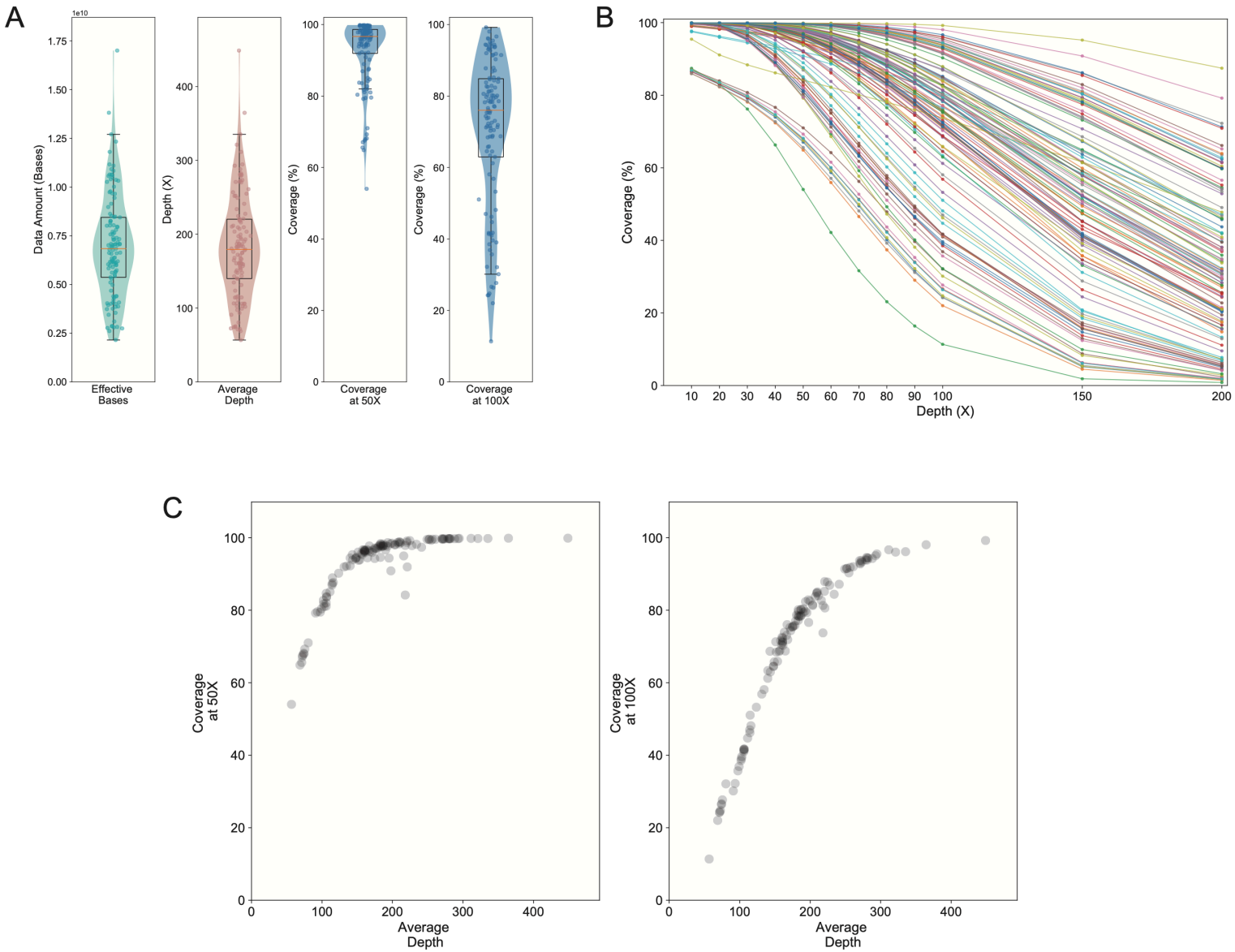

A
1e10
Data Amount (Bases)
Effective Bases
Depth (X)
Average Depth
Coverage (%)
Coverage at 50X
Coverage (%)
Coverage at 100X
B
Coverage (%)
Depth (X)
C
Coverage at 50X
Average Depth
Coverage at 100X
Average Depth

Figure S1. **Summary of exome sequencing data** (A) Distribution of effective sequenced bases, quantile-adjusted average depth, coverage of targeted region that covered at lease 50X, and coverage of targeted region that covered at least 100X. Each dot indicates a sequenced sample. (B) Distribution of exome coverage at specified minimum depth. (C) Sequencing saturation at 50X and 100X minimum depth.

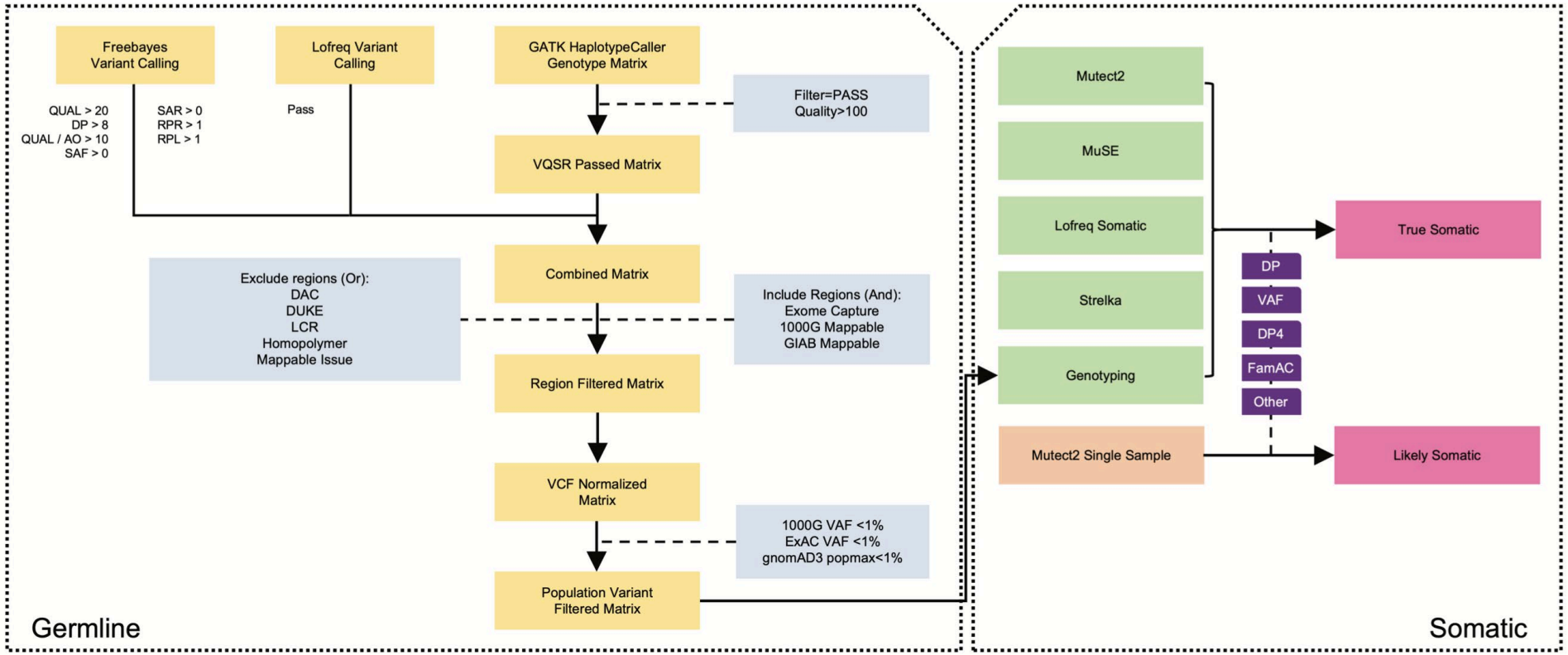
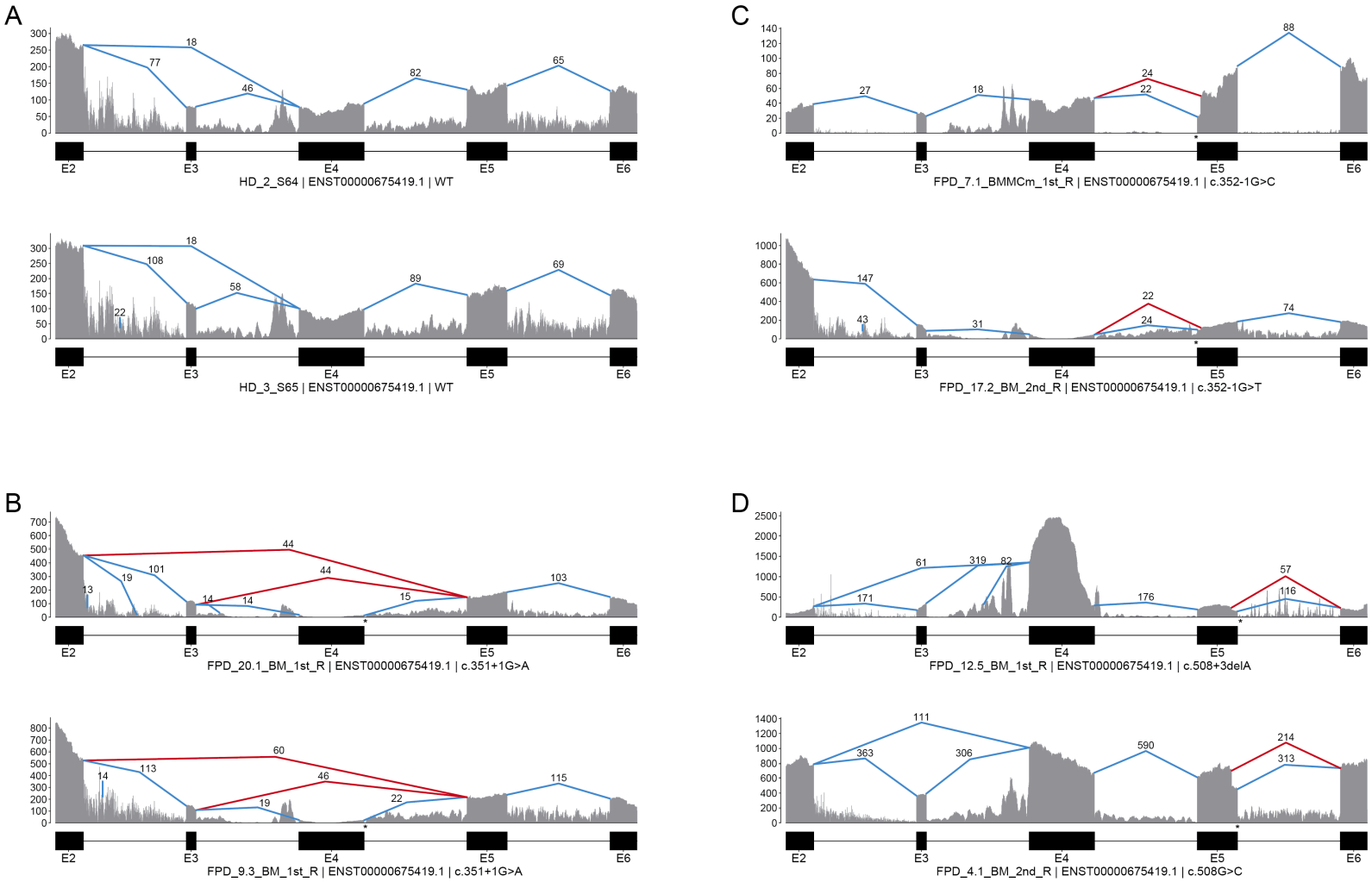


Figure S2. **Workflow of germline genotyping and somatic mutation calling**



Supplemental Figure 3. **Junction plots showing *RUNX1* splice junctions in two normal controls and 7 patients (in additional to those shown in Fig. 2A-G)**

Figure S4

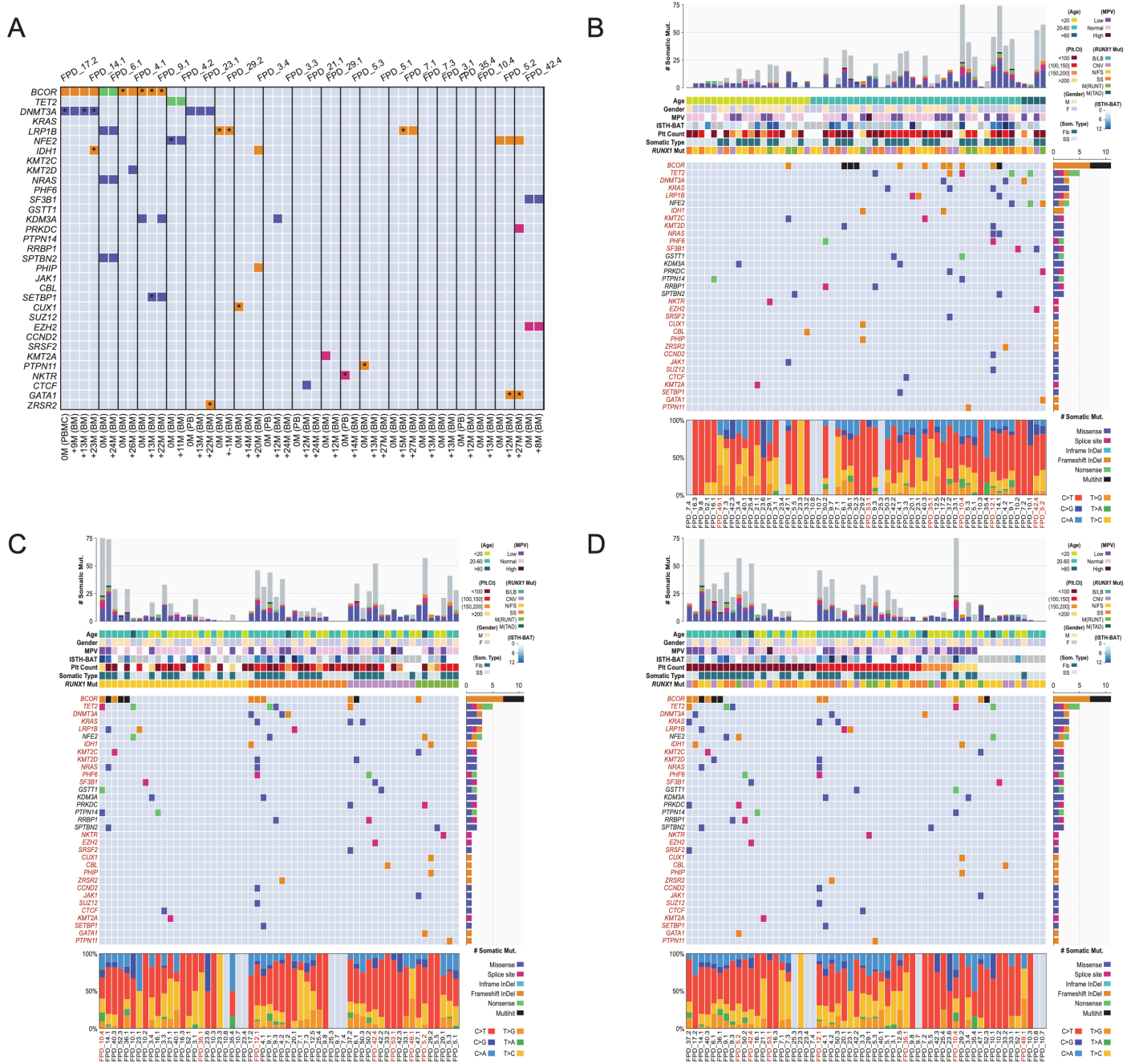
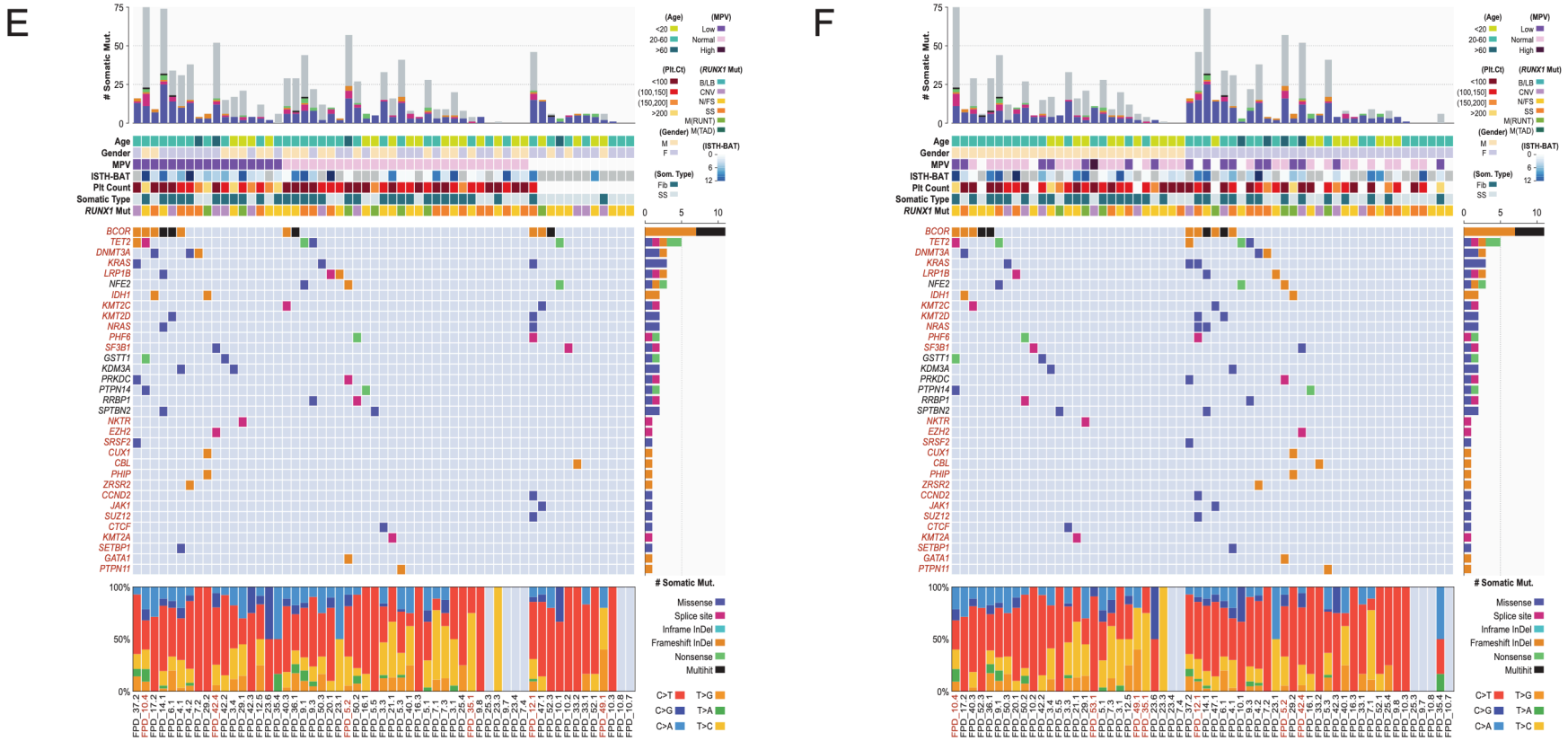
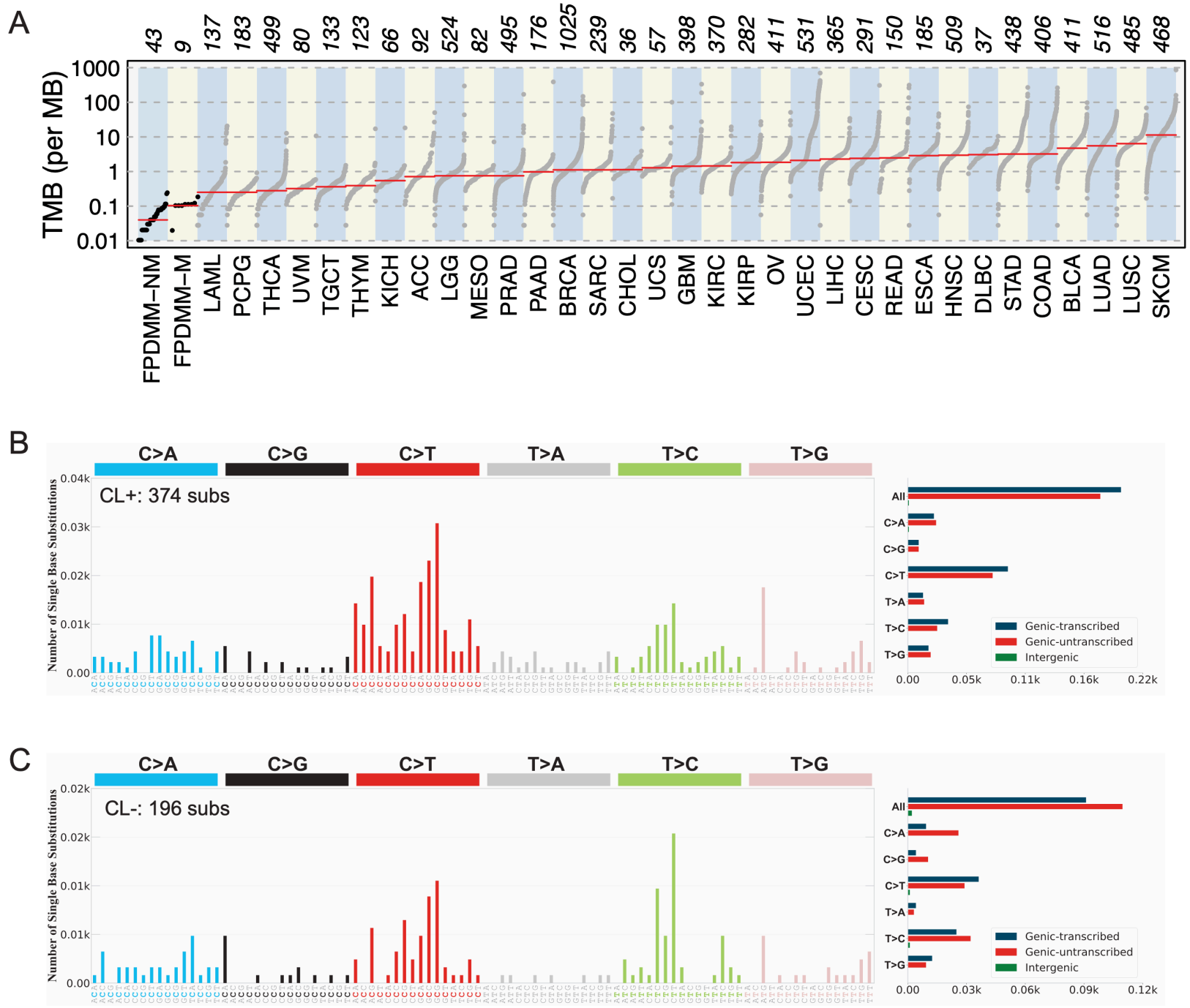

A
B
C
D
BCOR
TET2
DNMT3A
KRAS
LRP1B
NFE2
IDH1
KMT2C
KMT2D
NRAS
PHF6
SF3B1
GSTT1
KDM3A
PRKDC
PTPN14
RRBP1
SPTBN2
PHIP
JAK1
CBL
SETBP1
CUX1
SUZ12
EZH2
CCND2
SRSF2
KMT2A
PTPN11
NKTR
CTCF
GATA1
ZRSR2
Somatic Mut.
Missense
Splice site
Inframe InDel
Frameshift InDel
Nonsense
Multihit
C>T
C>G
C>A
T>G
T>A
T>C

Figure S4 (cont.)

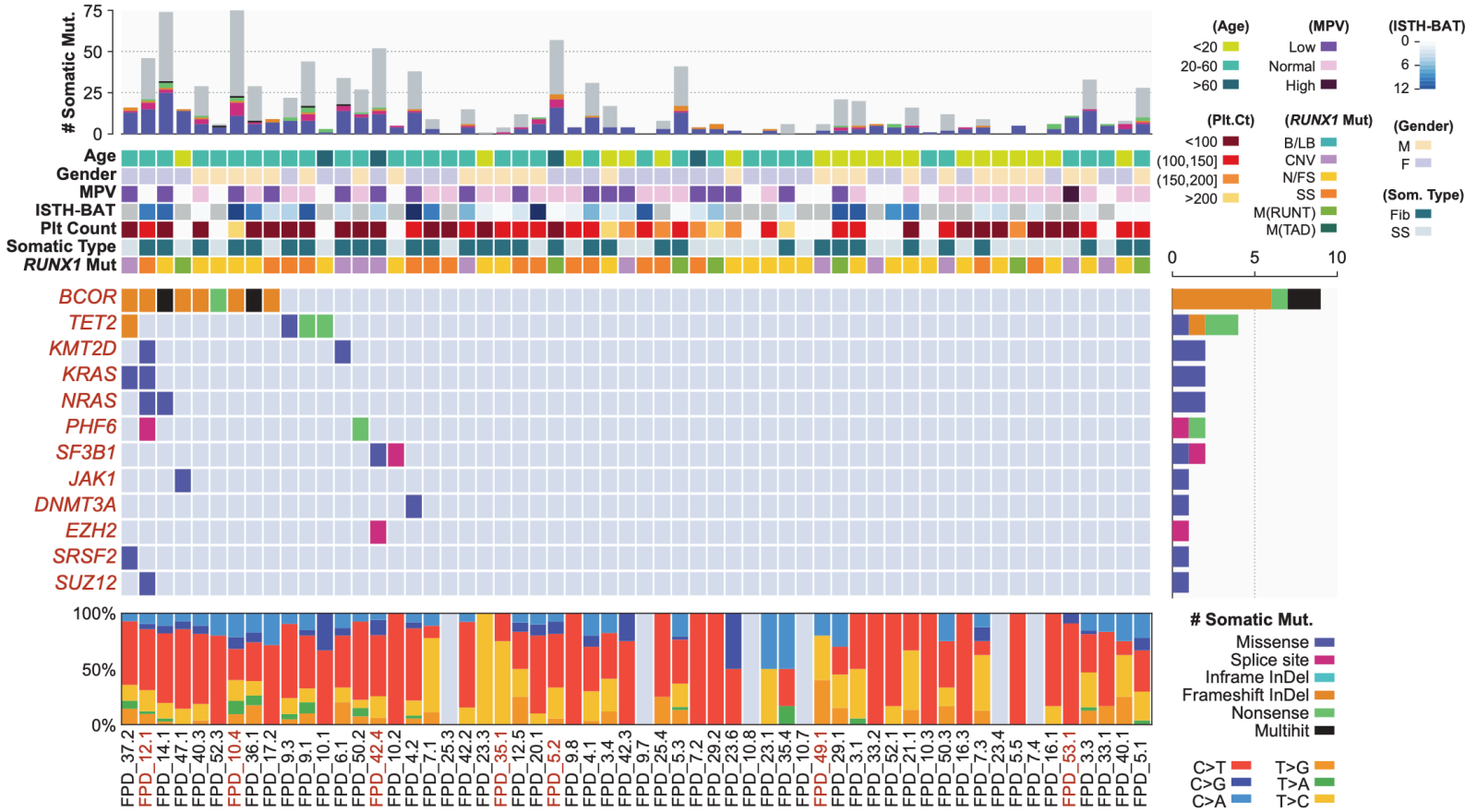


Supplemental Figure 4. **Somatic mutation comut plots related to Figure 3A** (A) Breakdown comut plot of patients with samples from multiple timepoints. Star signs indicate the mutation VAF<3%. Comut plot as shown in Fig 3A, being sorted by age (B), *RUNX1* mutation types (C), platelet count (D), mean platelet volume (MPV) (E) , and gender (F).

Figure S5

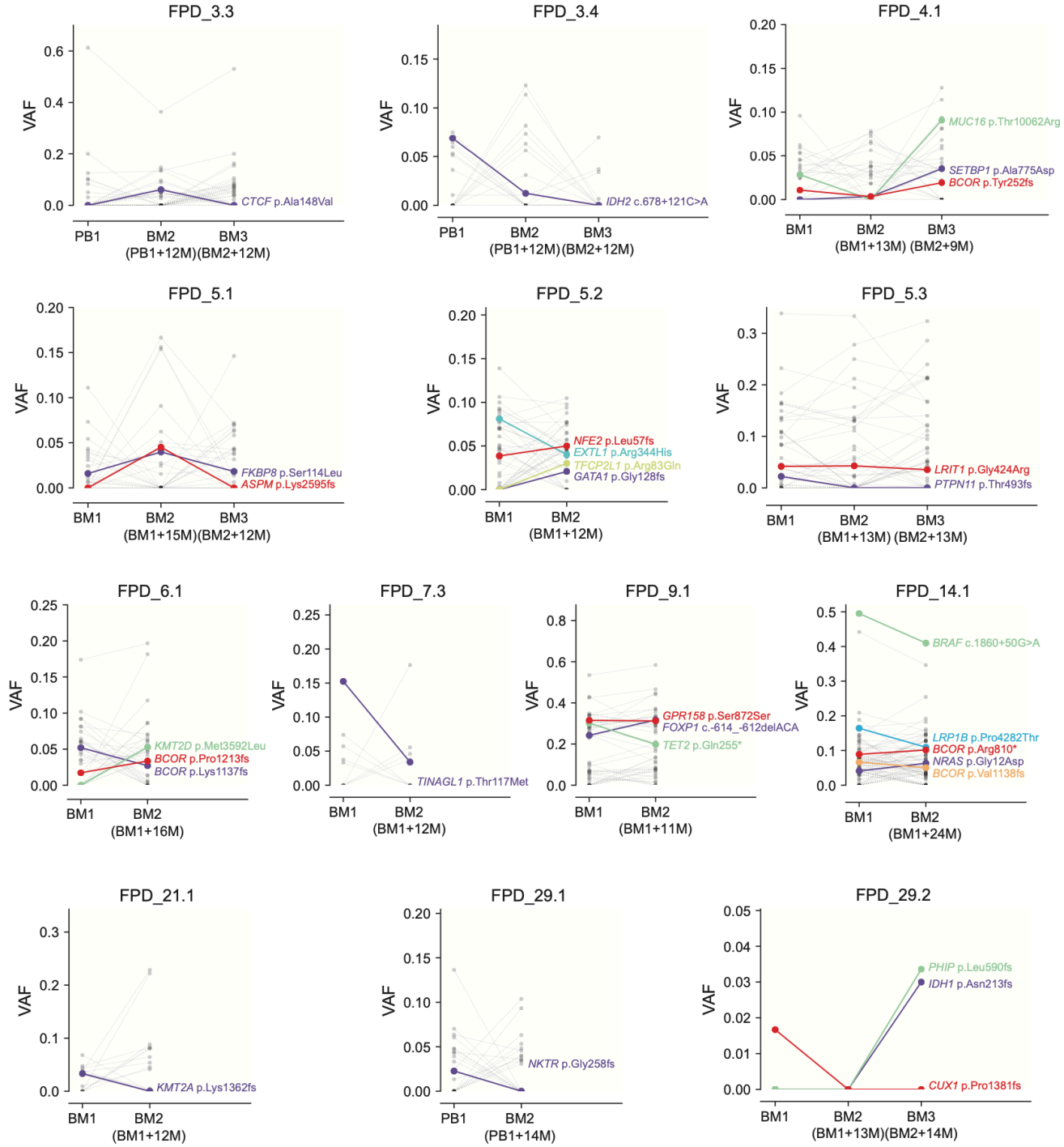

A
43 9 137 183 499 80 133 123 66 92 524 82 495 176 1025 239 36 57 398 370 282 411 531 365 291 150 185 509 37 438 406 411 516 485 468
TMB (per MB)
1000
100
10
1
0.1
0.01
FPDMM–NM
FPDMM–M
LAML
PCPG
THCA
UVM
TGCT
THYM
KICH
ACC
LGG
MESO
PRAD
PAAD
BRCA
SARC
CHOL
UCS
GBM
KIRC
KIRP
OV
UCEC
LIHC
CESC
READ
ESCA
HNSC
DLBC
STAD
COAD
BLCA
LUAD
LUSC
SKCM
B
C>A
C>G
C>T
T>A
T>C
T>G
CL+: 374 subs
Number of Single Base Substitutions
0.04k
0.03k
0.02k
0.01k
0.00
All
C>A
C>G
C>T
T>A
T>C
T>G
Genic-transcribed
Genic-untranscribed
Intergenic
0.00
0.05k
0.11k
0.16k
0.22k
C
C>A
C>G
C>T
T>A
T>C
T>G
CL-: 196 subs
Number of Single Base Substitutions
0.02k
0.02k
0.01k
0.01k
0.00
All
C>A
C>G
C>T
T>A
T>C
T>G
Genic-transcribed
Genic-untranscribed
Intergenic
0.00
0.03k
0.06k
0.09k
0.12k

Supplemental Figure 5. **Characterizations of somatic mutations in the NIH FPDMM cohort**. (A) Somatic mutation burdens of non-malignant, and malignant FPDMM patients, compared to the somatic mutation burdens of other cancer types from TCGA project. TMB: tumor mutational burden. Somatic mutation signatures of non-malignant FPDMM patients who carry somatic mutations in CHIP genes and AML driver genes (B) and of those who do not (C).



Supplemental Figure 6. **Somatic mutation comut plots of variants detected in 74 CHIP gene list and with allele frequency (VAF) >5%.**

Figure S7


FPD_3.3
VAF
CTCF p.Ala148Val
PB1
BM2
BM3
(PB1+12M)(BM2+12M)
FPD_3.4
IDH2 c.678+121C>A
(PB1+12M)(BM2+12M)
FPD_4.1
MUC16 p.Thr10062Arg
SETBP1 p.Ala775Asp
BCOR p.Tyr252fs
BM1
(BM1+13M) (BM2+9M)
FPD_5.1
FKBP8 p.Ser114Leu
ASPM p.Lys2595fs
(BM1+15M)(BM2+12M)
FPD_5.2
NFE2 p.Leu57fs
EXTL1 p.Arg344His
TFCP2L1 p.Arg83Gln
GATA1 p.Gly128fs
(BM1+12M)
FPD_5.3
LRIT1 p.Gly424Arg
PTPN11 p.Thr493fs
(BM1+13M) (BM2+13M)
FPD_6.1
KMT2D p.Met3592Leu
BCOR p.Pro1213fs
BCOR p.Lys1137fs
(BM1+16M)
FPD_7.3
TINAGL1 p.Thr117Met
(BM1+12M)
FPD_9.1
GPR158 p.Ser872Ser
FOXP1 c.-614_-612delACA
TET2 p.Gln255*
(BM1+11M)
FPD_14.1
BRAF c.1860+50G>A
LRP1B p.Pro4282Thr
BCOR p.Arg810*
NRAS p.Gly12Asp
BCOR p.Val1138fs
(BM1+24M)
FPD_21.1
KMT2A p.Lys1362fs
(BM1+12M)
FPD_29.1
NKTR p.Gly258fs
(PB1+14M)
FPD_29.2
PHIP p.Leu590fs
IDH1 p.Asn213fs
CUX1 p.Pro1381fs
(BM1+13M)(BM2+14M)

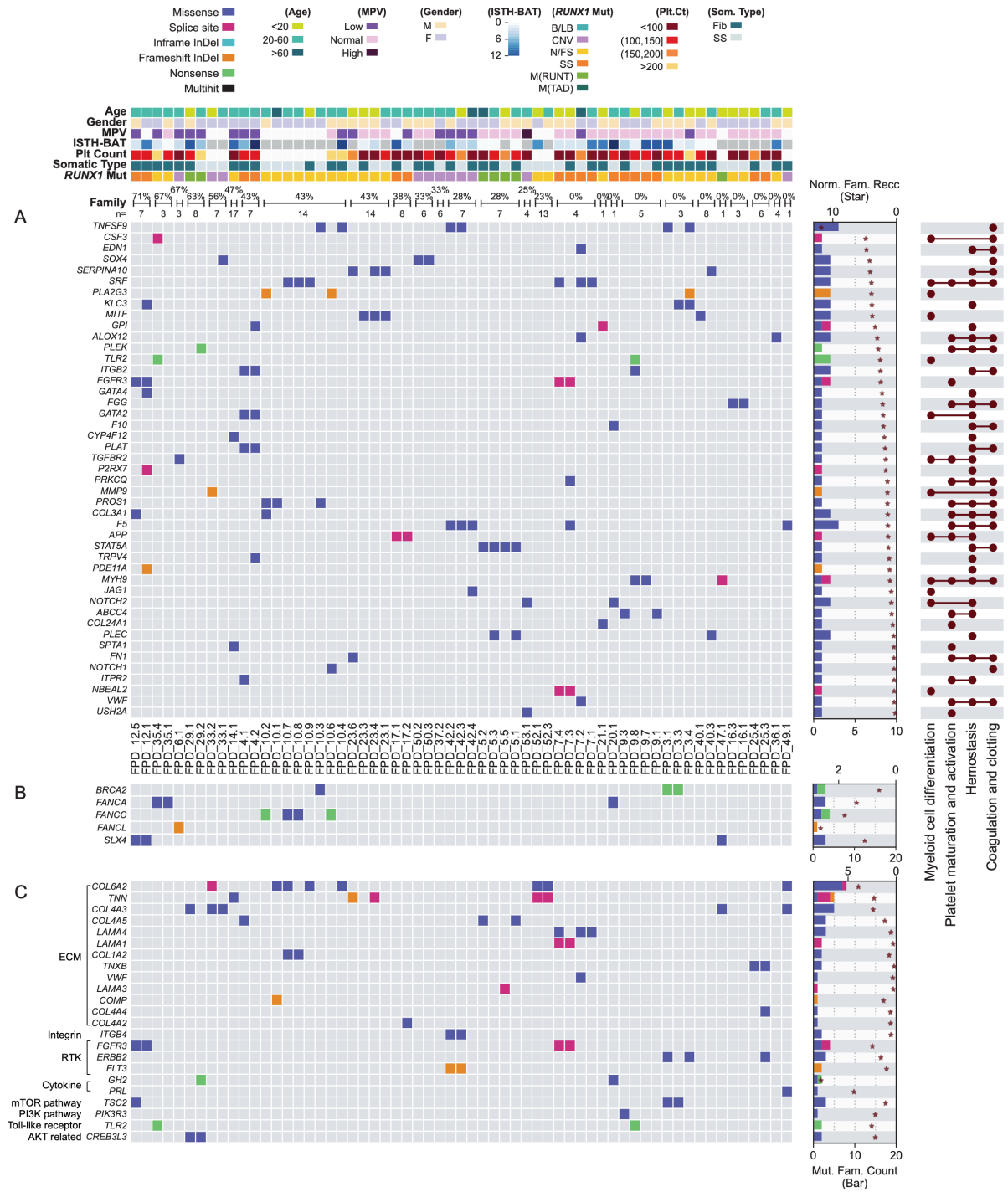
Supplemental Figure 7. **Somatic mutation VAF changes across samples collected in different timepoints from 13 non-malignant FPDMM patients.** Colored dots and lines show mutations in CL and leukemia genes. Grey dots and lines show somatic mutations in other genes.

Figure S8


A
FPD_7.3 | Age=6 | Platelet<100
Malig=False | Gender=m | RUNX1:splice site
4
1
4
FPD_7.3_BM_1st_E
FPD_7.3_BM_2nd_E
FPD_3.4 | Age=6 | Platelet>=150
Malig=False | Gender=m | RUNX1:nonsense
FPD_3.4_PB_1st_E
FPD_3.4_BM_2nd_E
6
1
7
0
1
0
2
FPD_3.4_BM_3rd_E
FPD_21.1 | Age=8 | Platelet<100
Malig=False | Gender=m | RUNX1:frameshift
6
4
6
FPD_21.1_BM_2nd_E
FPD_21.1_BM_1st_E
FPD_29.1 | Age=10 | 100<=Platelet<150
Malig=False | Gender=m | RUNX1:missense
9
4
8
FPD_29.1_PB_1st_E
FPD_29.1_BM_2nd_E
FPD_3.1 | Age=12 | 100<=Platelet<150
Malig=False | Gender=m | RUNX1:nonsense
11
2
7
FPD_3.1_PB_1st_E
FPD_3.1_BM_2nd_E
FPD_7.1 | Age=30 | Platelet<100
Malig=False | Gender=f | RUNX1:splice site
4
1
4
FPD_7.1_BM_2nd_E
FPD_7.1_BM_1st_E
FPD_6.1 | Age=31 | Platelet<100
Malig=False | Gender=f | RUNX1:deletion large
8
16
10
FPD_6.1_BM_1st_E
FPD_6.1_BM_3rd_E
FPD_4.1 | Age=37 | 100<=Platelet<150
Malig=False | Gender=f | RUNX1:splice site
FPD_4.1_BM_1st_E
FPD_4.1_BM_2nd_E
2
3
3
11
4
1
7
FPD_4.1_BM_3rd_E
FPD_3.3 | Age=39 | 100<=Platelet<150
Malig=False | Gender=m | RUNX1:nonsense
FPD_3.3_PB_1st_E
FPD_3.3_BM_2nd_E
3
1
9
1
3
2
14
FPD_3.3_BM_3rd_E
FPD_5.3 | Age=51 | 100<=Platelet<150
Malig=False | Gender=f | RUNX1:missense
FPD_5.3_BM_1st_E
FPD_5.3_BM_2nd_E
8
2
2
19
0
2
8
FPD_5.3_BM_3rd_E
FPD_5.1 | Age=52 | 100<=Platelet<150
Malig=False | Gender=m | RUNX1:missense
FPD_5.1_BM_2nd_E
FPD_5.1_BM_1st_E
6
1
5
5
5
3
8
FPD_5.1_BM_3rd_E
FPD_35.4 | Age=54 | Platelet>=150
Malig=False | Gender=f | RUNX1:nonsense
0
2
4
FPD_35.4_BM_1st_E
FPD_35.4_BM_2nd_E
FPD_10.4 | Age=58 | Platelet>=150
Malig=True | Gender=m | RUNX1:nonsense
47
30
3
FPD_10.4_BM_2nd_E
FPD_10.4_BM_1st_E
FPD_4.2 | Age=59 | Platelet>=150
Malig=False | Gender=f | RUNX1:splice site
FPD_4.2_PB_1st_E
FPD_4.2_BM_2nd_E
1
1
5
3
16
12
0
FPD_4.2_BM_3rd_E
FPD_14.1 | Age=60 | Platelet<100
Malig=False | Gender=f | RUNX1:frameshift
10
55
9
FPD_14.1_BM_1st_E
FPD_14.1_BM_2nd_E
FPD_9.1 | Age=61 | Platelet<100
Malig=False | Gender=m | RUNX1:splice site
3
34
7
FPD_9.1_BM_1st_E
FPD_9.1_BM_2nd_E
FPD_42.4 | Age=72 | Platelet<100
Malig=True | Gender=f | RUNX1:deletion large
5
36
11
FPD_42.4_BM_1st_E
FPD_42.4_BM_In-between_E
B
Malignancy
False
True
Somatic Mutation
Age
Platelet Count (K/mcL)
Platelet<100
100<=platelet<150
Platelet>=150
Somatic Mutation
Age
Gender
M
F
Somatic Mutation
Age

Supplemental Figure 8. **Somatic mutations detected at more than one timepoints.** (A) Venn plots of somatic mutations identified at different timepoints. Numbers indicate mutations detected at one or more timepoints. (B) Numbers of somatic mutations correlating with patient age, malignancy status, platelet count, and gender. Each dot represent one patient's sample.

Figure S9



Supplemental Figure 9. **Germline variants in genes related to (A) Myeloid cell differentiation; Platelet maturation and activation; Hemostasis; Coagulation and clotting, (B) Fanconi anemia, and (C) RTK-RAS-PI3K pathway.** The design and color codes for this figure are similar to those in Fig 3A, S5, and S6.