Supplemental material

Table of contents

Supplemental Methods pages 2-13

Supplemental Figure S1. Germline variant status of related donors. page 14

Supplemental Figure S2. Spectrum of identified presumed germline variants. page 15

Supplemental Methods

Data source

The Center for International Blood and Marrow Transplant Research (CIBMTR) is a nonprofit research collaboration of the National Marrow Donor Program (NMDP)/The Match and the Medical College of Wisconsin (MCW). More than 300 medical centers worldwide submit clinical data to the CIBMTR about hematopoietic stem cell transplantation (HSCT) and other cellular therapies. Participating centers are required to report all transplantations consecutively. The CIBMTR ensures data quality through computerized checks for discrepancies, physicians' review of submitted data, and on-site audits of participating centers. The CIBMTR complies with federal regulations that protect human research participants. The Institutional Review Boards of MCW, NMDP and The University of Chicago approved this study.

DNA extraction

DNA was extracted from previously frozen (-80^{o}C) peripheral blood that was collected from HSCT recipients prior to initiation of transplant conditioning chemotherapy and from their related donors. DNA extraction was performed according to standard protocols using a QIAGEN QIAcube automated nucleic acid extraction instrument with a sample loading volume of 200µL. Double stranded DNA was quantified via fluoroscopy using an Invitrogen Qubit Fluorometer. A second DNA extraction was performed for samples yielding a low DNA concentration (<5 ng/µL).

Augmented whole exome sequencing

Whole exome sequencing (WES) was augmented with custom designed spike-in probes to cover non-coding regions known to contain inherited risk alleles including: the 5'UTR of *ANKRD26* (NM_014915), the 5'UTR of *DKC1* (NM_001363), intron 31 of *FANCI* (NM_ 001113378). WES was performed at the Yale Center for Genome Analysis using 500 ng of genomic DNA, where available. For samples with a genomic DNA concentration <9 ng/µL (by fluoroscopy), the quantity of DNA used reflected the maximum input sample volume of 55 µL. The lowest quantity of DNA used was 97.5 ng.

Genomic DNA was sheared to a mean fragment length of approximately 220 bp using focused acoustic energy (Covaris E220). Fragmented sample size distribution was determined by using the Caliper LabChip GX system. Fragmented DNA samples are transferred to a 96-well plate and library construction is completed using a liquid handling robot. Following fragmentation, T4 DNA polymerase and T4 polynucleotide kinase created blunt end and phosphorylated the fragments. The large Klenow fragment then added a single adenine residue to the 3' end of each fragment and custom adapters (IDT) were ligated using T4 DNA ligase. Magnetic AMPure XP beads (Beckman Coulter) were utilized to purify and size select the adapter-ligated DNA fragments. The adapter-ligated DNA fragments were then PCR amplified using custom-made primers (IDT). During PCR, a unique 10 base index was inserted at both ends of each DNA fragment. Sample concentration was determined by picogreen and inserts size distribution by using the Caliper LabChip GX system. Samples yielding at least 1 ug of amplified DNA were used for capture.

Equal amounts of each sample were pooled prior to capture. The dried sample was reconstituted according to the manufacturer's protocol (IDT), heat-denatured, and mixed with biotinylated DNA probes produced by IDT (xGen Exome Panel). Hybridizations were performed at 65°C for 16 hours. After capture the samples were

mixed with streptavidin-coated beads and washed with a series of stringent buffers to remove non-specifically bound DNA fragments. The captured fragments were PCR amplified and purified with AMPure XP beads. Samples were quantified by quantitative real time PCR (RT-qPCR) using a commercially available kit (KAPA Biosystems) and insert size distribution were determined with the LabChip GX. All 799 samples submitted for augmented WES had a yield of ≥0.5 ng/μl and were used for sequencing.

Prepared library concentrations were normalized to 2nM and loaded onto Illumina NovaSeq6000 S4 flow cells at a concentration yielding at least 750 gigabases of passing filter data per lane (*i.e.* > 3 terabases for S4 flow-cell). Loading concentration for Exome libraries have been optimized at 400pM to maximize both well occupancy and unique read output while limiting duplicates associated with patterned flow cell technology and ExAmp chemistry utilized on Novaseqs. Samples were sequenced using 101 bp paired-end sequencing reads according to Illumina protocols. Dual 10 bp indices were read during additional sequencing reads that automatically follow the completion of read.

Signal intensities were converted to individual base calls during a run using the system's Real Time Analysis (RTA) software. Base calls were transferred from the machine's dedicated personal computer to the Yale High Performance Computing cluster via a 1 Gigabit network mount for downstream analysis. Primary analysis, sample de-multiplexing and alignment to the human genome, was performed using Illumina's CASAVA 1.8.2 software suite. The sample error rate was less than 2% and the distribution of reads per sample in a lane was within reasonable tolerance. The average coverage for the 799 samples sequenced was 40X.

Variant annotation and calling

An in-house custom GATK 3.7 pipeline was used to call single nucleotide variants (SNVs) and small insertions and deletions. Fastq files were converted to unmapped BAM (uBAM) files and read group information was added using FastqToSam (Picard 2.8.1). MarkIlluminaAdapters (Picard 2.8.1) was used to mark the 5‘ start position of the adapter sequences. Reads were aligned to the human reference genome (GRCh38) with BWA 0.7.15 and merged with uBAM files using MergeBamAlignment (Picard 2.8.1) to create mapped BAM files. Duplicates were marked. Base quality score recalibration (BQSR) was conducted using the dbSNP database to build a model of covariation based on a set of known variants. BQSR was applied to the sequencing data to generate BAM files with accurate base substitution, insertion, and deletion quality scores prior to variant calling and generation of VCF files. Variants in U2AF1, *PRPF8* and *SMARCB1* were initially not called due to issues with the alignment to GRCh38, a common problem that has recently been published in regard to *U2AF1* variant calling.[1] Based on the published suggestions, we were able to adjust our pipeline accordingly.

Variants were annotated to include information about gene, cDNA and amino acid changes, population allele frequency from the Genome Aggregation Database (gnomAD v2.1.1)[2] and the Exome Aggregation Consortium (ExAC),[3] ClinVar variant interpretation,[4] genotype, allele depth, sequence depth, and genotype quality. Variants with less than 10 total reads; with a population frequency greater than or equal to 0.005 in gnomAD v2.1.1; and/or non-splicing intronic variants, except for intronic variants in genes known to harbor recurrent deleterious intronic variants including genes such as *GATA2, FANCI, BRCA1, BRCA2, IKZF1*, and *NF1,* were excluded from manual variant curation.

Genomic copy number variants (CNVs) were called using the ExomeDepth R package (version 1.1.12), which we optimized by using a reference set of 22

unrelated, sex-matched samples. The ExomeDepth package[5] utilized this reference set to identify genomic regions in each patient sample that contained significantly increased or reduced read depths in order to identify CNVs.

Variant interpretation

233 genes associated with inherited hematopoietic malignancies, BMF syndromes, TBD, DNA repair deficiency, immunodeficiency, RASopathies, general tumor predisposition syndromes and congenital cytopenias were analyzed. For germline variant interpretation, recommendations of the 2015 American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP),[6] the criteria for sequence variant interpretation from the Clinical Genome Sequence Variant Interpretation Working Group[7–15] and practical recommendations[16] were followed. Specified ACMG/AMP rules established by ClinGen variant curation expert panels were used for analysis of *ITGA2B/ITGB3*,[17] *PTEN*,[18] RASopathies,[19] *RUNX1*[20–22] and *TP53*.[23] The variant analysis includes the type of variant (e.g. missense, nonsense, frameshift), the location within the protein and potential functional domains/variant hotspots, prediction of nonsense-mediated decay and splicing, previously reported variants, *in vivo* and *in vitro* functional studies, presence of the variant in population databases, computational *in silico* ensemble predictions, and association with the mechanism of disease and the underlying phenotype.

Somatic variants in all 404 recipients and 25 donors sharing a germline variant were assessed in an additional 167 genes known to be somatically mutated in MDS based on large MDS datasets from cBioPortal[24] and the Catalogue Of Somatic Mutations In Cancer.[25] For somatic variants, recommendations of the 2017

ACMG/AMP ACMG/AMP and American Society of Clinical Oncology guidelines were followed.[26]

Presumed germline status was determined for P/LP variants that were not shared with a VAF within germline range (0.4-0.6 or 1), no other potential P/LP variants within germline range in the same recipient, and a combination of the following criteria based on a multiple-criteria decision analysis: **(1)** presence and frequency of confirmed somatic variants (nonsynonymous, in-frame indels and truncating variants only) in the gene of question in patients with MDS, therapy-related MDS or chronic myelomonocytic leukemia (no somatic variants reported [6], somatic variants in less than 1% [2] or more than 1% of patients [0]); **(2)** previous report of the identical variant as a somatic variant, which is not appliclabe for truncating variants (unless reported) since they are rarely recurrent (not reported [2], ≤5 occurences [1], hotspot > 10 occurences [0]); **(3)** additional somatic variants in other genes and their VAF (none [5], none at the same VAF ± 15% [3], one or more at the same VAF ± 15% [0]); **(4)** reported predisposing condition (yes [7], no [0]); **(5)** MDS subtype with higher bone marrow blast count (>5%) or advanced disease (no [3], yes [0]); and **(6)** high pretest probability as has been described for genes such as *BRCA1/2, DDX41, GATA2, RUNX1* and others[27–30] (yes [1], no [0]). A variant was considered presumed germline when a score of ≥6 was reached.

Variant validation

Sanger sequencing

All pathogenic (P)/likely pathogenic (LP) germline SNVs were confirmed by Sanger sequencing in the recipient and the donor (if shared).

Subcloning

Compound heterozygous variants were confirmed to be *in trans* by subcloning using the NEB PCR cloning kit (E1202S, New England Biolabs) and consecutive Sanger sequencing. *SBDS* and its pseudogene *SBDSP1* share 97% homology. In order to avoid mismapping to the pseudogene, sequences upstream to the variant were used to discriminate the presence of a true *SBDS* variant from a contaminating *SBDSP1* sequence.

RT-qPCR

All potential CNVs derived from the bioinformatic analysis were evaluated by RT-q-PCR. Primers were designed and used to amplify patient DNA for specific exons within the amplified/deleted CNV region using the StepOnePlus Real-Time PCR System (Applied Biosystems) in triplicate. PCR products were analyzed with StepOne Software v2.3 (Applied Biosystems). One no template control was included per primer pair per run. Amplification efficiency was calculated for each run by using serial dilutions. The standard 2-DDCT method was used to calculate copy number ratio.

Statistical analysis

Overall survival (OS) was defined as the time from transplant to death from any cause. Disease free survival (DFS) was defined as survival following HSCT without relapse or progression. Patients who survived without evidence of disease relapse or progression were censored at last follow-up. Non-relapse mortality (NRM) was defined as death from any cause before one-month post-HSCT or death in continuous remission. Primary graft failure was defined as no evidence of engraftment of donor cells within the first month after transplant without evidence of disease relapse. Acute and chronic graft versus host disease (GvHD) were defined and graded as per the standard criteria.[31,32] Cytogenetics were subclassified into

favorable, intermediate and poor categories based on the International Prognostic Scoring System.[33]

The Kaplan-Meier estimator was used to calculate probabilities of OS and DFS with variance estimated by Greenwood's formula. The cumulative incidence of relapse, acute GvHD, chronic GvHD, NRM, neutrophil engraftment, platelet recovery, and graft failure were calculated using the cumulative incidence estimator while accounting for competing risks. Comparison of survival and cumulative incidence curves was done using the log-rank test and Gray's test, respectively. Estimates of outcomes were reported as probabilities with 95% confidence intervals. Given the small sample size, a multivariate analysis was not performed. Univariable outcomes were analyzed for the recipients by recipient and donor P/LP variant status (no P/LP variant vs. healthy carriers vs. shared variant). Healthy carriers refer to recipients who received donor stem cells lacking an autosomal dominant P/LP variant and/or cells with a single P/LP allele with an autosomal recessive mode of inheritance. Shared variant refers to recipients who received donor stem cells that shared the same autosomal dominant P/LP variant.

Comparison of somatic patterns across different groups was carried out by odds ratio analysis.

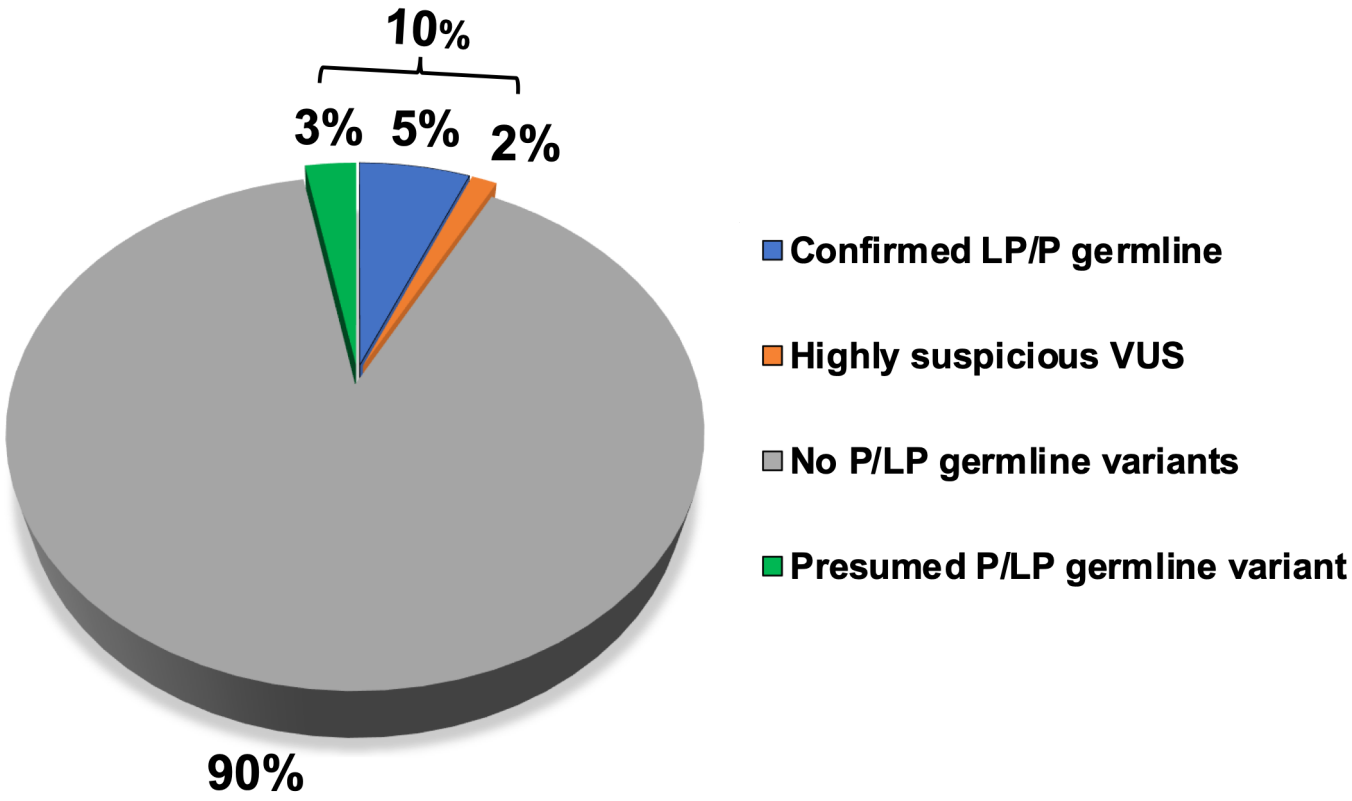
References

1. Miller CA, Walker JR, Jensen TL, et al. Failure to Detect Mutations in U2AF1 due to Changes in the GRCh38 Reference Sequence. *J. Mol. Diagnostics*. 2022;24(3):219–223.
2. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434–443.
3. Lek M, Karczewski KJ, Minikel E V., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285–291.
4. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 2016;44(D1):D862–D868.
5. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*. 2012;28(21):2747–2754.
6. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 2015;17(5):405–24.
7. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum. Mutat.* 2018;39(11):1517–1524.
8. Biesecker LG, Harrison SM. The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genet. Med.* 2018;20(12):1687–1688.
9. Brnich SE, Rivera-Muñoz EA, Berg JS. Quantifying the potential of functional evidence to reclassify variants of uncertain significance in the categorical and Bayesian interpretation frameworks. *Hum. Mutat.* 2018;39(11):1531–1541.

10. Brnich SE, Abou Tayoun AN, Couch FJ, et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* 2020;12(1):3.
11. Ghosh R, Harrison SM, Rehm HL, Plon SE, Biesecker LG. Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Hum. Mutat.* 2018;39(11):1525–1530.
12. Riggs ER, Andersen EF, Cherry AM, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* 2020;22(2):245–257.
13. Strande NT, Brnich SE, Roman TS, Berg JS. Navigating the nuances of clinical sequence variant interpretation in Mendelian disease. *Genet. Med.* 2018;20(9):918–926.
14. Tavtigian S V., Greenblatt MS, Harrison SM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* 2018;20(9):1054–1060.
15. Walsh MF, Ritter DI, Kesserwan C, et al. Integrating somatic variant data and biomarkers for germline variant classification in cancer predisposition genes. *Hum. Mutat.* 2018;39(11):1542–1552.
16. Feurstein S, Hahn CN, Mehta N, Godley LA. A practical guide to interpreting germline variants that drive hematopoietic malignancies, bone marrow failure, and chronic cytopenias. *Genet. Med.* 2022;1–24.
17. Ross JE, Zhang BM, Lee K, et al. Specifications of the variant curation guidelines for ITGA2B / ITGB3 : ClinGen Platelet Disorder Variant Curation Panel. *Blood Adv.* 2021;5(2):414–431.

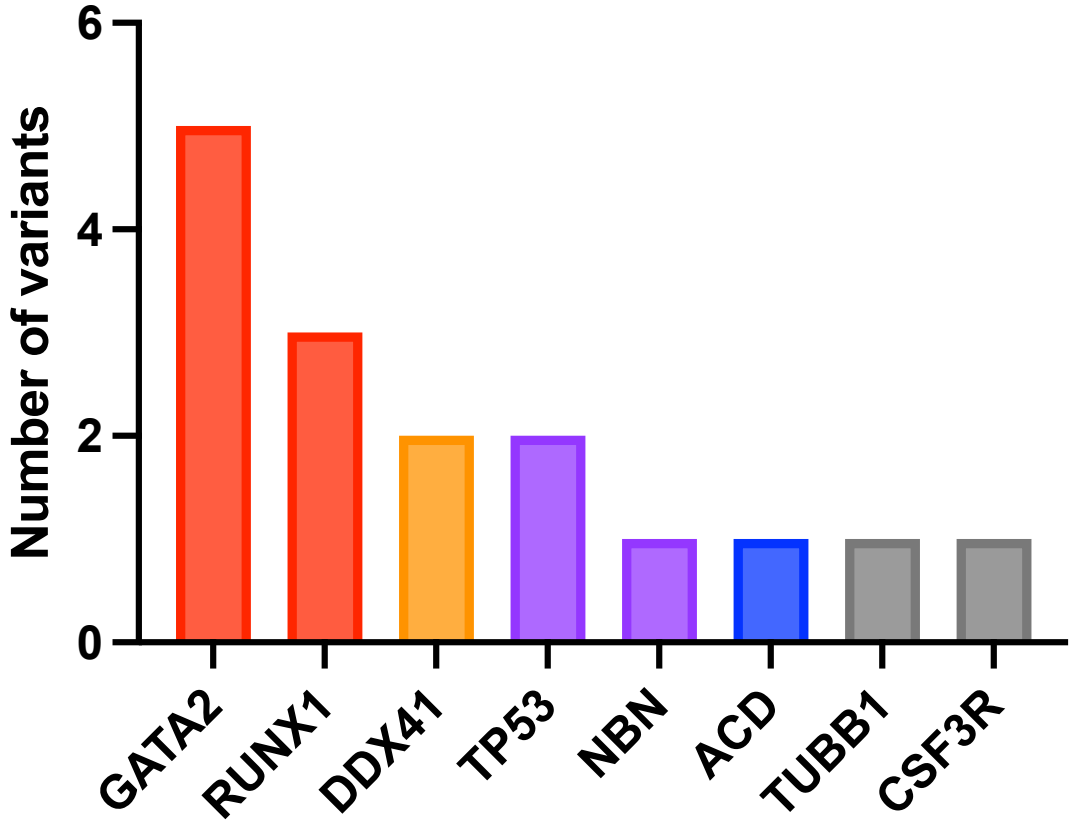
18. Mester JL, Ghosh R, Pesaran T, et al. Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. *Hum. Mutat.* 2018;39(11):1581–1592.
19. Gelb BD, Cavé H, Dillon MW, et al. ClinGen's RASopathy Expert Panel consensus methods for variant interpretation. *Genet. Med.* 2018;20(11):1334–1345.
20. Luo X, Feurstein S, Mohan S, et al. ClinGen Myeloid Malignancy Variant Curation Expert Panel recommendations for germline RUNX1 variants. *Blood Adv.* 2019;3(20):2962–2979.
21. Wu D, Luo X, Feurstein S, et al. How I curate: applying American Society of Hematology-Clinical Genome Resource Myeloid Malignancy Variant Curation Expert Panel rules for RUNX1 variant curation for germline predisposition to myeloid malignancies. *Haematologica*. 2020;105(4):870–887.
22. Feurstein S, Zhang L, DiNardo CD. Accurate germline RUNX1 variant interpretation and its clinical significance. *Blood Adv.* 2020;4(24):6199–6203.
23. Fortuno C, Lee K, Olivier M, et al. Specifications of the ACMG/AMP variant interpretation guidelines for germline TP53 variants. *Hum. Mutat.* 2020;(December):humu.24152.
24. Cerami E, Gao J, Dogrusoz U, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discov.* 2012;2(5):401–404.
25. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 2019;47(D1):D941–D947.
26. Li MM, Datto M, Duncavage EJ, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. *J. Mol. Diagnostics*. 2017;19(1):4–23.

27. Bannon SA, Routbort MJ, Montalban-Bravo G, et al. Next-Generation Sequencing of DDX41 in Myeloid Neoplasms Leads to Increased Detection of Germline Alterations. *Front. Oncol.* 2021;10(January):1–6.
28. Feurstein S, Drazer M, Godley LA. Germline predisposition to hematopoietic malignancies. *Hum. Mol. Genet.* 2021;30(R2):R225–R235.
29. Kraft IL, Godley LA. Identifying potential germline variants from sequencing hematopoietic malignancies. *Hematology*. 2020;2020(1):219–227.
30. Simon L, Spinella J-F, Yao C-Y, et al. High frequency of germline RUNX1 mutations in patients with RUNX1-mutated AML. *Blood*. 2020;135(21):1882–1886.
31. Shulman HM, Sullivan KM, Weiden PL, et al. Chronic graft-versus-host syndrome in man. A long-term clinicopathologic study of 20 Seattle patients. *Am. J. Med.* 1980;69(2):204–17.
32. Przepiorka D, Weisdorf D, Martin P, et al. 1994 Consensus Conference on Acute GVHD Grading. *Bone Marrow Transplant.* 1995;15(6):825–8.
33. Greenberg P, Cox C, LeBeau MM, et al. International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood*. 1997;89(6):2079–2088.



Supplemental Figure S1. Germline variant status of related donors. This pie chart shows that confirmed LP/P germline variants in genes with autosomal dominant inheritance were shared in 5% of the donors. Presumed germline variants that were not shared with the recipient (based on a VAF within germline range and a multiple-criteria decision analysis) were found in 3% of donors. Two percent of donors shared a highly suspicious VUS with their recipient (based on calculating posterior probabilities in a Bayesian framework).

Abbreviations: LP, likely pathogenic; P, pathogenic; VAF, variant allele frequency; VUS, variant of unknown significance.



Supplemental Figure S2. Spectrum of identified presumed germline variants. This bar chart shows the number of presumed LP/P germline variants in eight different genes (*GATA2, RUNX, DDX41, TP53, NBN, ACD, TUBB1, CSF3R*). Colors indicate gene groups associated with transcription (red), RNA helicase (orange), DNA repair (purple), TBD (blue) and others (grey). Abbreviations: LP- likely pathogenic, P-pathogenic, TBD- telomere biology disorder.