Detailed Methods

Sample Processing:

To establish the guiding principles for germline tissues, a prospective study was conducted to assess somatic call rates of WES data generated for four candidate germline tissues including biopsies of normal skin, eyebrow hair follicles, ≥95% pure CD3+ T-cells, and buccal cells relative to bone marrow from 16 histologically confirmed MDS cases. Bone marrow mononuclear cells (BMMNCs) were isolated from bone marrow aspirates. Whole bone marrow was passed through a blunt 22-gauge needle three times and treated twice with Pharm Lyse solution for RBC lysis (Becton Dickinson) and then washed 2x with PBS. The cells were flash frozen in liquid nitrogen and stored in DNase/RNase free microfuge tubes at -80˚C until DNA extraction. Controls that underwent pilot feasibility testing included eyebrow hairs (12 per patient, 6 collected per brow), a skin biopsy (2-4mm), skin swab (x2 swabs swiped 10 times on forearm), urine (50 ml), 10 nail clippings per patient, and T cells (CD3+). Details on T cell purification are defined below. Buccal swabs were added after study initiation on 11 patients with a collection procedure that was optimized to recover epithelial cells (supplemental material). Buccal swab and eyebrow hair follicle DNA was extracted using the QIAamp DNA Investigator Kit (Qiagen, Hilden, Germany, Cat. No.56504) according to the manufacturer’s protocol. Briefly, ATL buffer with proteinase-K was added to the sample, placed on a thermomixer, and incubated at 56°C with shaking at 900 rpm for 2 hours. Lysates from all samples were stored in a -80ºC freezer until isolation. The DNA was then purified using minElute columns and quantified via Qubit fluorometric quantitation (QubitTM 3.0 Fluorometer, ThermoFisher, Scientific, Waltham, MA). DNA quality (DNA Integrity Number, DIN) was determined using an Illumina TapeStation (Agilent

Genomics, Santa Clara, CA). Human DNA content was determined from buccal swabs using the TaqMan copy number reference assay for measurement of human Telomere Reverse Transcriptase (hTERT) (Applied Biosystems, Foster City, CA) (data not shown). Due to overt leukocyte contamination detected by the Multistix urine test strip (Roche Diagnostics, Indianapolis, IN) in urine collections, low DNA quality of nail clippings, and insufficient DNA quantity from skin swabs, these samples were deemed poor candidates for NGS germline tissue and therefore discontinued from the study (supplemental material).

Buccal cell collection:

Saliva contains a mixture of epithelial cells and leukocytes. Procedures were tested and optimized for the collection of epithelial cells. Several types of swabs were first tested to define the optimal swab for cell removal. Patients were given two sterile cotton swabs (Puritan Medical Products, 258051PCTTFABUSA, Guilford, ME) and placed in a plastic transport tube. The procedure was first explained to the participant and then the patient rinsed for 20 seconds with water and expectorated the rinse. One cotton swab was removed from the sterile culturette tube and used to swab the inner cheek 20 times rotating to cover all surfaces of the cotton tip. This was then allowed to air dry for one minute prior to placing the swab back into the plastic culturette container for overnight transport. Two swabs were collected from each patient. Both swabs were agitated in 30 ml PBS to release epithelial cells. The cells were then centrifuged at 1,000 g for 10 minutes followed by aspiration of the PBS prior to lysis (as defined above).

T-cell isolation:

T cells were purified from the peripheral blood mononuclear cells (PBMCs) by negative selection using the EasySep Human T Cell Enrichment Kit (Stem Cell Technologies,19051, Vancouver, Canada) and then subjected to a second round of purification or flow cytometry sorting to generate populations with >95% purity. T cell purity was assessed using flow cytometry. Cells were labeled with CD3-FITC antibody (Stem Cell Technologies, 60011FI.1) and Propidium Iodidie (Miltenyi Biotec, Bergisch Gladbach, Germany) and acquired on a LSRII cytometer (Becton Dickinson, Franklin Lakes, NJ).

***DNA quantification and quality assessment**:*

DNA was quantified using QubitTM 3.0 Fluorometer, ThermoFisher, Scientific, Waltham, MA with DNA quality (DNA Integrity Number, DIN) determined using an Illumina TapeStation (Agilent Genomics, Santa Clara, CA).

Whole Exome Sequencing:

WES was performed on up to four candidate controls (normal skin obtained during the bone marrow procedure, eyebrow hair follicles, purified CD3+ T-cells, and buccal swabs and bone marrow mononuclear cells (BM-MNCs) for each case (as detailed on page 3 of the main manuscript). 1.5 ug of DNA was used to perform SureSelect exome enrichment (Agilent Genomics) libraries and sequenced on a HiSeq2500 Sequencer (Illumina Inc, San Diego, CA) with the goal of 100x mean on-target coverage (Figure S1).

Targeted Resequencing:

Targeted resequencing was achieved using a custom NimbleGen Hybrid Capture System (Roche Sequencing, Pleasanton, CA) to validate all putative candidate variants identified by WES using MuTect and/or Strelka as variant callers (see settings below). Only those variants identified via both WES and targeted validations were considered somatic. Using the Nextera DNA library preparation kit (Illumina Inc, San Diego, CA), eight barcoded libraries were pooled for each enrichment pull down. The target sequencing depth was 400x.

Whole genome amplification (WGA):

WGA was performed using the Illustra GenomiPhi V2 DNA amplification kit (GE Healthcare Life Sciences, Pittsburgh, PA).

Alignment, Variant calling and Settings:

FASTQ files were aligned to human reference hs37d5 using BWA 0.7.7 [PMC2705234] in paired-end mode. Alignments are refined with PICARD 1.82 (http://picard.sourceforge.net/) MarkDuplicates and GATK Lite 2.2-16 [PMC3083463] RealignerTargetCreator/IndelRealigner and BaseRecalibrator.

Sequence reads from exome and targeted validation sequencing were aligned with the Burrows-Wheeler Aligner (BWA)[1] to the human genome v37 (hs37d5) and somatic variants called for each BM-MNC:germline pair using both Strelka and MuTect[2,3] Putative somatic mutations were defined as those that were observed by both Strelka and MuTect or listed as PASS in Strelka (Figure S1). NimbleGen Hybrid Capture System (Roche Sequencing, Pleasanton, CA) was used to validate putative nucleotide variants detected by Strelka/MuTect software using native (non-WGA) DNA from the bone marrow and control tissues. Somatic mutations were identified from the WES and target capture validation sequencing data with a combination of Strelka 1.0.13[3] (default exome settings, ssnvNoise = 0.00000005, sindelNoise = 0.0000001) and MuTect 1.1.4[2] (--max_alt_alleles_in_normal_count 3, --max_alt_allele_in_normal_fraction 0.05; indels detected with GATK Lite 2.2-16 SomaticIndelDetector). Custom perl code and VCFtools

0.1.12b[4] were used to convert and integrate Strelka and MuTect output to VCF format. Mutations were considers "passing" if observed as (PASS in Strelka) OR (observed at all in Strelka AND PASS in MuTect). Mutations were additionally detected in the target capture validation data using VarScan 2.4.3[5] using suggested settings from the developer page and DREAM-3 settings for FPFILTER. Varscan somatic: --min-coverage 3 --min-var-freq 0.08 --p-value 0.10 --somatic-p-value 0.05 --strand-filter 0 --output-vcf 1. Varscan fpfilter: --dream3-settings 1. Bam-readcount 0.8.0 was used to generate the readcount file required for fpfilter: -q 1 -b 20. Mutations were annotated using ANNOVAR[6], 1000 Genomes, the Exome Sequencing Project, COSMIC, and ClinVar. Data were analyzed, visualized and reviewed using perl, R, IGV[7], VarSifter[8], and Circos[9].

For an added quality control measure, a subset of common single nucleotide polymorphisms (SNPs) (>15% VAF in 1000 Genomes, Figure S3) were assessed to ensure that the bone marrow and germline samples were derived from the same individual. Briefly, all genomic variants (including germline) were detected in each sample with GATK. Variant positions observed in the 1000 Genomes project at >15% VAF were compared across all samples to calculate concordance. All-vs.-all samples concordances were plotted in a heatmap style output. Initially, an inappropriate matching was observed between the MDS01 Tumor and all of the MDS16 samples (Figure S3A). Following investigation and correction of this issue, the resulting concordance plot shows no inappropriate matching (Figure S4B).

Supplemental Tables and Figures

Table S1. Baseline Characteristics of Myelodysplastic Syndrome Patients

Mean Age (range), n=26	69.7 (48-82)
Male:Female ratio, n=26	17 (65%): 9 (35%)
Mean disease duration in months (range), n=26	22.9 (0.08 – 51.6)
0-6 months (n,%)	3 (11.5)
7-18 months (n, %)	7 (26.9)
19-36 months (n,%)	11 (42.4)
>37 months (n, %)	5 (19.2)

WHO classification	**Diagnosis (n=26*, all) (n, %)**	**Diagnosis (n=16**, sequenced) (n, %)**
RA	1 (3.8)	1 (6.3)
RARS	4 (15.4)	4 (25)
RCMD/RCMD-RS	8 (30.7)	4 (25)
5q Syndrome	0	0
RAEBI/II	6 (23.0)	5 (31)
AML	5 (19.2)	1 (6.3)
CMML	0	0
MDS Unclassified	2 (7.8)	1 (6.3)
MDS/MPN Overlap	0	0

Abbreviations: MDS, myelodysplastic syndrome; RA, refractory anemia; RCMD, refractory cytopenia with multilineage dysplasia; WHO, World Health Organization. RCMD with multilineage dysplasia including patients with or without ringed sideroblasts; AML, acute myeloid leukemia; MPN, myeloproliferative neoplasm, RA with excess blasts −1 and 2. In this study, WES was omitted for patients with T-cell purity < 95% (n=5, MDS11, MDS12, MDS13, MDS16, and MDS22) as they were unavailable for repeated sample collection. MDS-14 failed to provide adequate germline control samples (n=1), and AML diagnosis after completion of the bone marrow staging resulted in withdrawal of

MDS07, MDS09, MDS10, and MDS19 (n=4), Therefore, final sequencing data is reported on 16 patients.

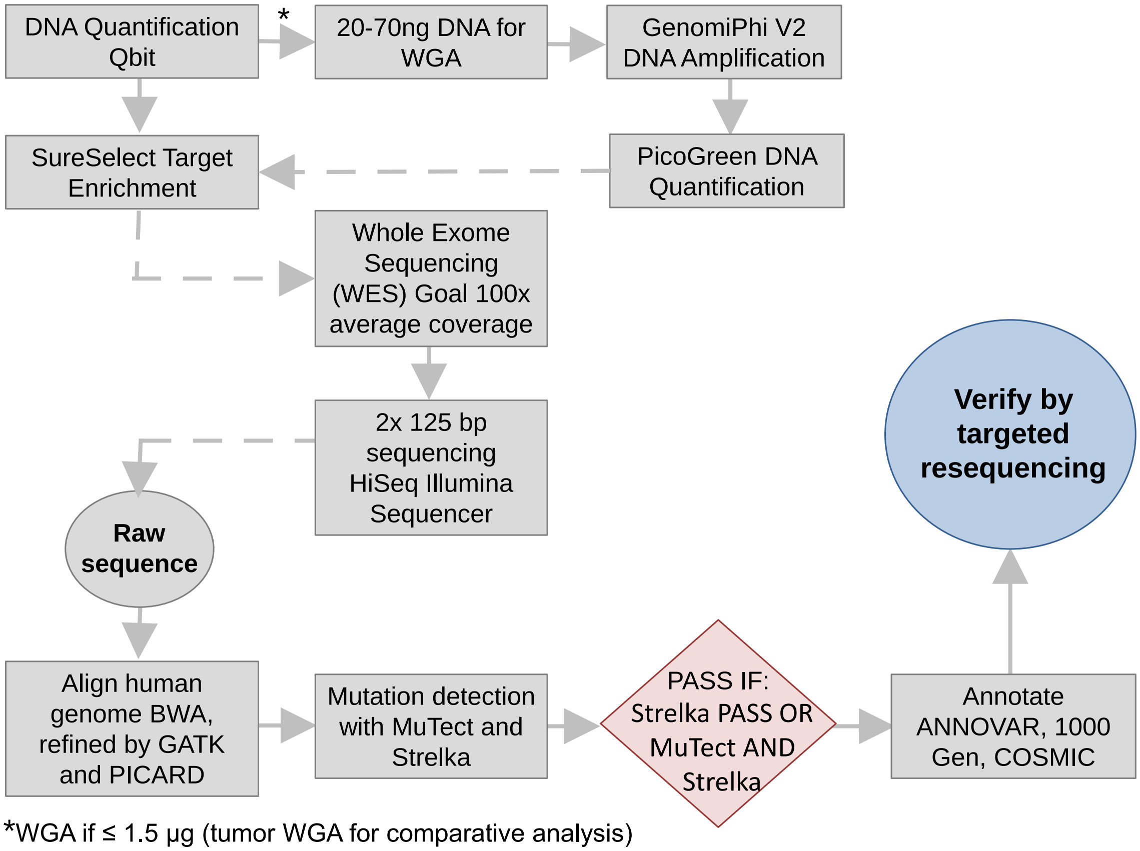
Table S2. Quality and quantity of DNA.

Tissue	**DNA quantity**		**DNA quality**	
	Mean [µg]	SD	Mean [DIN]	SD
Skin biopsy*	1	0.5	7.6	0.5
Eyebrow Hair	0.2	0.2	7.3	0.3
T-cells	4.7	1.9	8.8	0.5
Buccal swab	2.1	1.8	4.5	0.9
Bone Marrow	6.8	3.5	9	0.3
Urine**	3.2	3.2	5.3	1.9
Nail clippings	0.9	na	1.6	na
Skin swab #	<0.01	na	na	na

*DNA yield from 2mm biopsies (n=11) 0.7µg (SD 0.3);
4mm biopsies (n=5) 1.7 µg (SD 0.4)

**Range for females 14-200 ng/ml; range for males 4-60 ng/ml, urine discontinued due to excessive hematopoietic cell contamination with measurement by Acon Laboratories, Inc test kit for leukocyte contamination and then confirmation by cytospin of urine followed by H&E staining. Difficulty with the future shipment of liquids for The National MDS Natural History Study also drove our decision to proceed with sequencing buccal samples.

#Due to low sample quantities, quality was not assessed.


DNA Quantification Qbit
*
20-70ng DNA for WGA
GenomiPhi V2 DNA Amplification
SureSelect Target Enrichment
PicoGreen DNA Quantification
Whole Exome Sequencing (WES) Goal 100x average coverage
2x 125 bp sequencing HiSeq Illumina Sequencer
Raw sequence
Verify by targeted resequencing
Align human genome BWA, refined by GATK and PICARD
Mutation detection with MuTect and Strelka
PASS IF: Strelka PASS OR MuTect AND Strelka
Annotate ANNOVAR, 1000 Gen, COSMIC

*WGA if ≤ 1.5 µg (tumor WGA for comparative analysis)

Figure S1. *Workflow for sequencing and analysis.* Sequencing method and workflow for whole exome sequencing (WES), decision to perform whole genome amplification, and bioinformatics pipeline to identify variants for targeted resequencing.

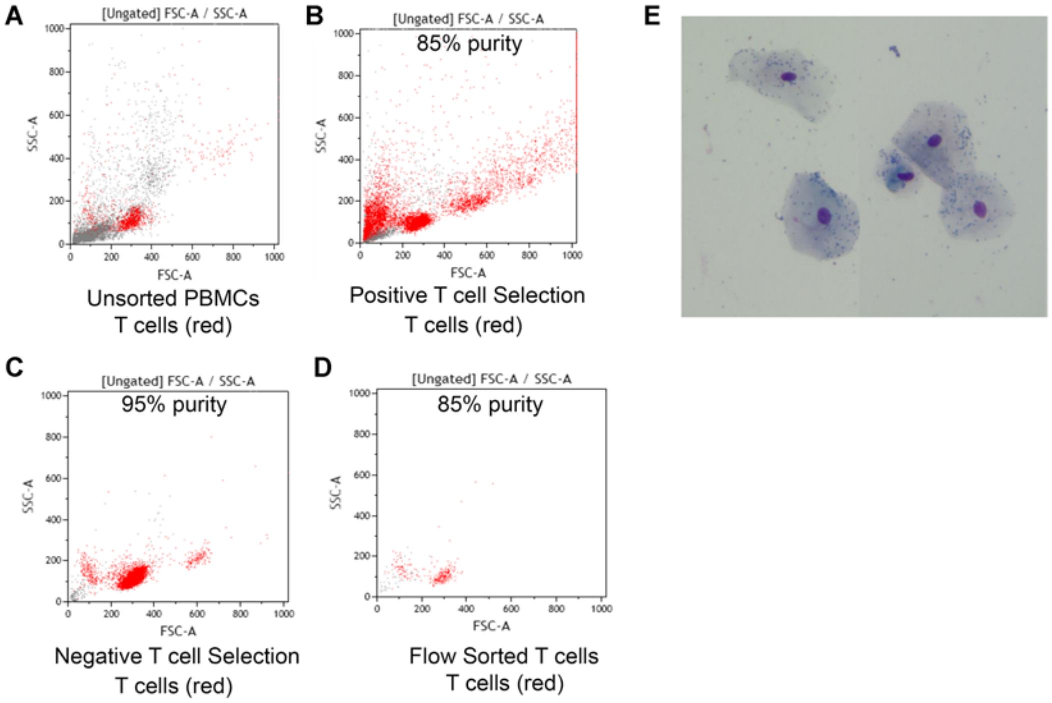


Figure S2. ***Flow cytometry for T-cells isolated by different methods and buccal cells.*** **(A)** unsorted peripheral blood, **(B)** All T-cells were isolated from peripheral blood mononuclear cells collected at the same time as the bone marrow aspiration. T-cells isolated by positive selection showing 85% purity but enhanced cellular debris (lower left corner) and/or changes in distribution of the forward and side scatter plot which is consistent with increased activation-associated cell death, **(C)** T-cells isolated by negative selection with 95% purity and little or no activation-associated cell death, and **(D)** flow sorted T-cells with 85% purity. This is a representative plot of each method. Data on purity was generated for each sample and described in Table S2. WES and targeted resequencing was limited to samples with ≥ 95% purity by negative selection of flow sorting. Positively selected samples were not used in this study for

sequencing. **(E)** Epithelial cells evident in buccal swab collection by hematoxylin & eosin (H&E) staining (20x magnification).

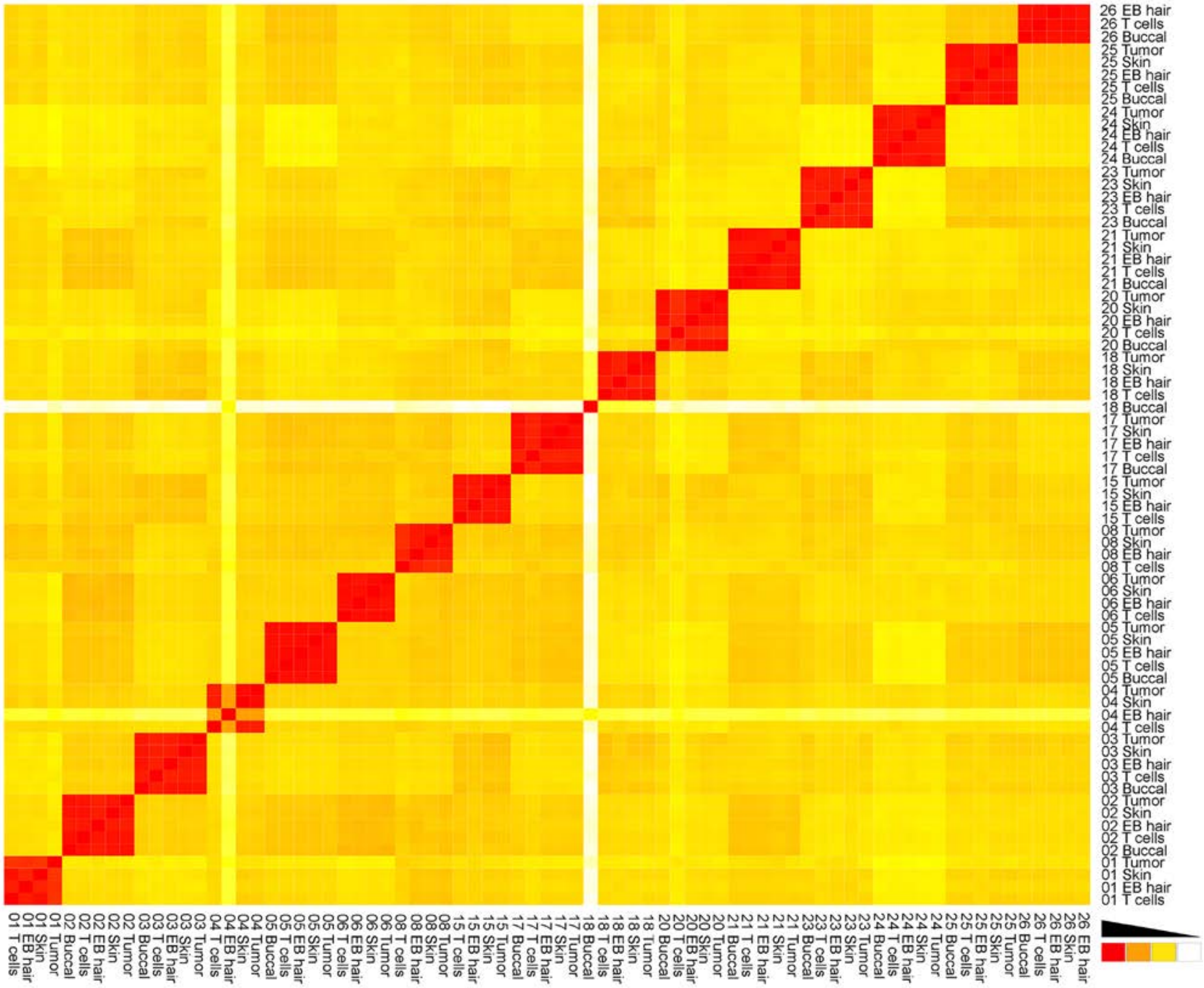


Figure S3. ***Quality control measures***. A subset of common single nucleotide polymorphisms (SNPs) (>15% VAF in 1000 Genomes) were assessed in germline samples that were derived from the same individual. All-vs.-all samples concordances were plotted in a heatmap style output. Inappropriate matching was observed between the MDS01 Tumor upon initial analysis. After identification of a sample swap, the resulting concordance plot shows no inappropriate matching after correction.

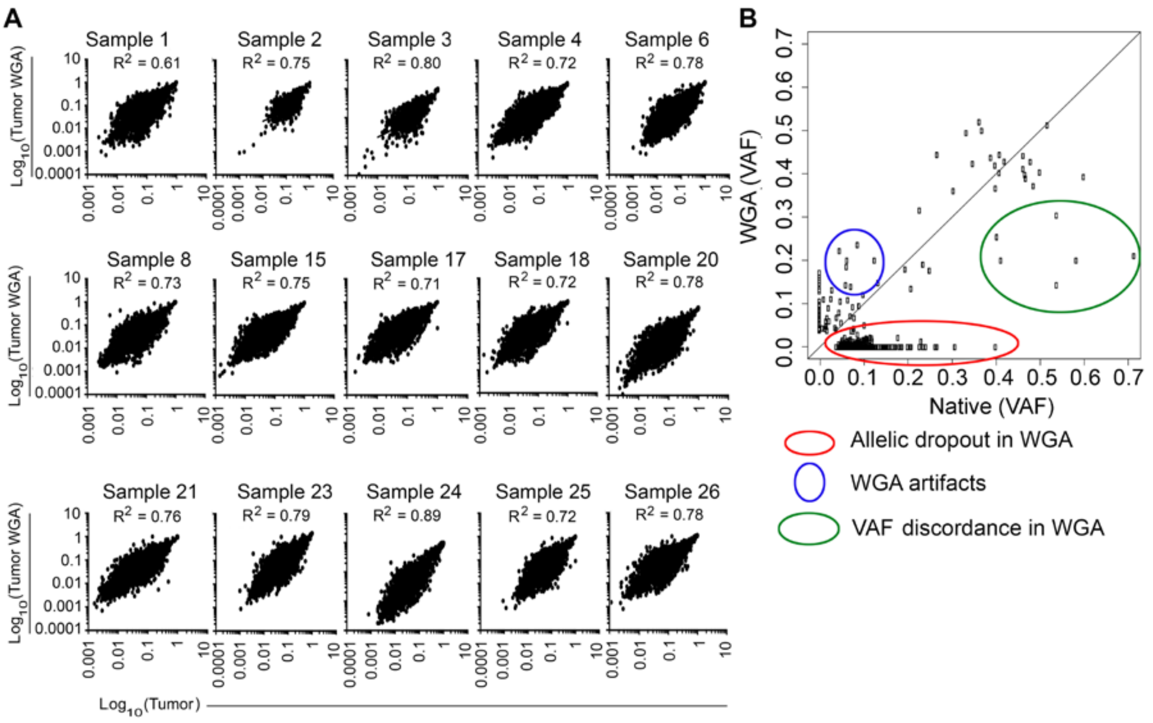


Figure S4. *Whole exome sequencing for Whole Genome Amplified (WGA) versus native DNA.* We show that higher DNA quantities are beneficial to improve coverage, prevent allele dropout, artifacts and reduced variant allele fraction (VAF) observed with whole genome amplification (WGA). **(A)** Scatter plot of native leukemia versus WGA leukemia samples for 15 of 16 patients. MDS05 resulted in poor WGA amplification and was thus excluded from this analysis, but WES was performed and analyzed as indicated. **(B)** Representative scatter plot of one patient demonstrating allele dropout, discordant variant allele fraction (VAF), and WGA artifacts.

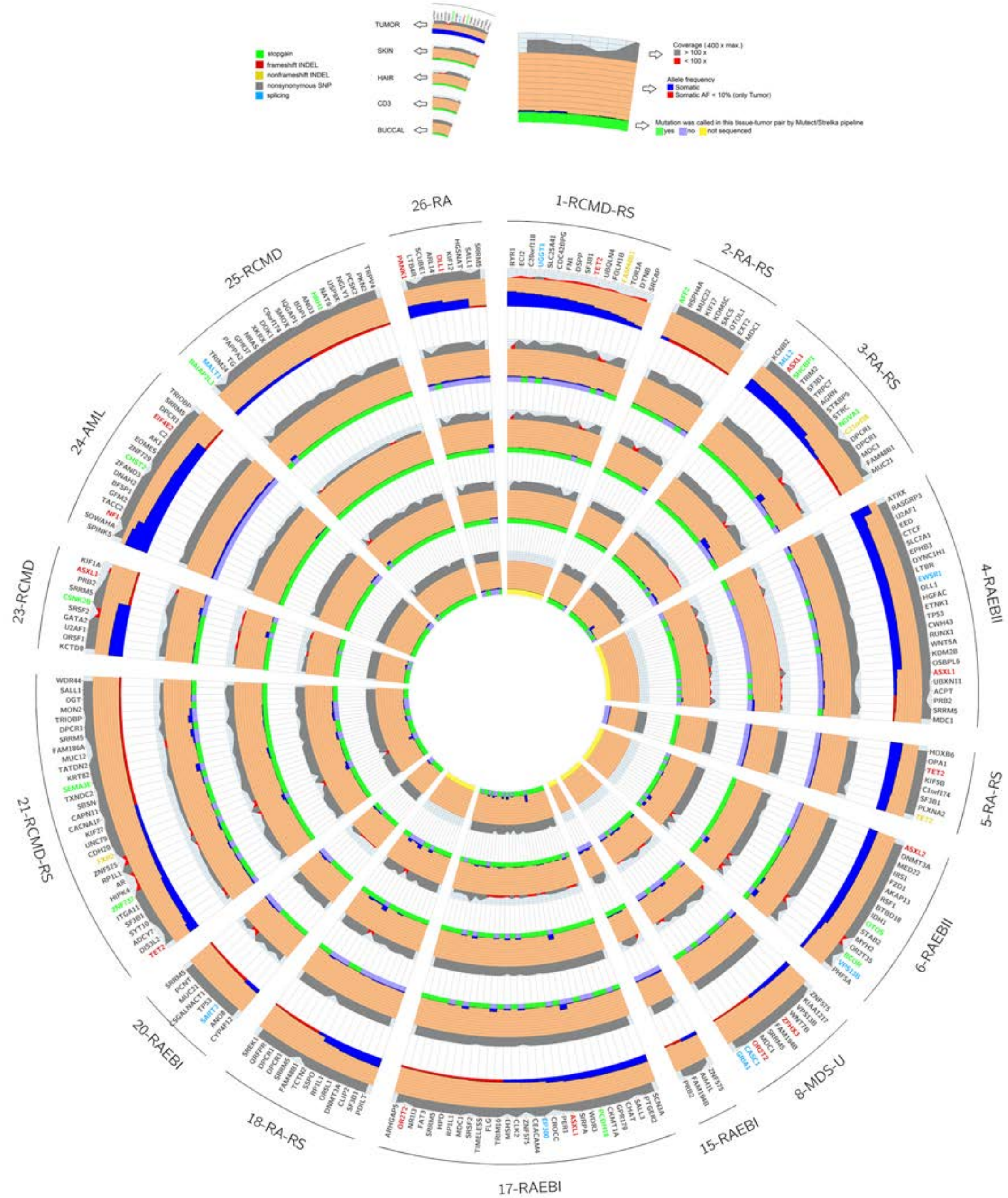


Figure S5. *Circos diagram of tumor germline sequencing metrics.* This Circos diagram was generated as described previously[9] and shows a representation of somatic mutations that occur in > 5% of the bone marrow. The five wedges represent four groups of

mutations called by the MuTect/Strelka[3] pipeline and a group with missing values due to lack of buccal swab sample collection. The gene names orthogonally to the radius represent protein changing somatic mutations, color coded by the kind of mutation. The five main rings represent the different tissues and are further separated in three sections. The outer section represents the coverage at the position of the somatic mutation, followed by the allelic fraction and the inner section represents the status of the four germline samples called by the MuTect/Strelka pipeline. Coverage, reference and variant allele fractions (VAFs), and predicted protein consequence (ANNOVAR)[6] across all capture-based resequenced variants by patient sample. Each ring represent a tissue type sequenced and those tumor-germline pairs which called the plotted variant are denoted green (gray if not called). Here, somatic variants were as expected in MDS including genes involved in RNA splicing *(SF3B1*), DNA methylation (*TET2, DNMT3A*) and chromatin modification (*ASXL1*). Skin biopsies showed higher degrees of neoplastic contamination, resulting in missed variant identification (varying based on bioinformatics methods).

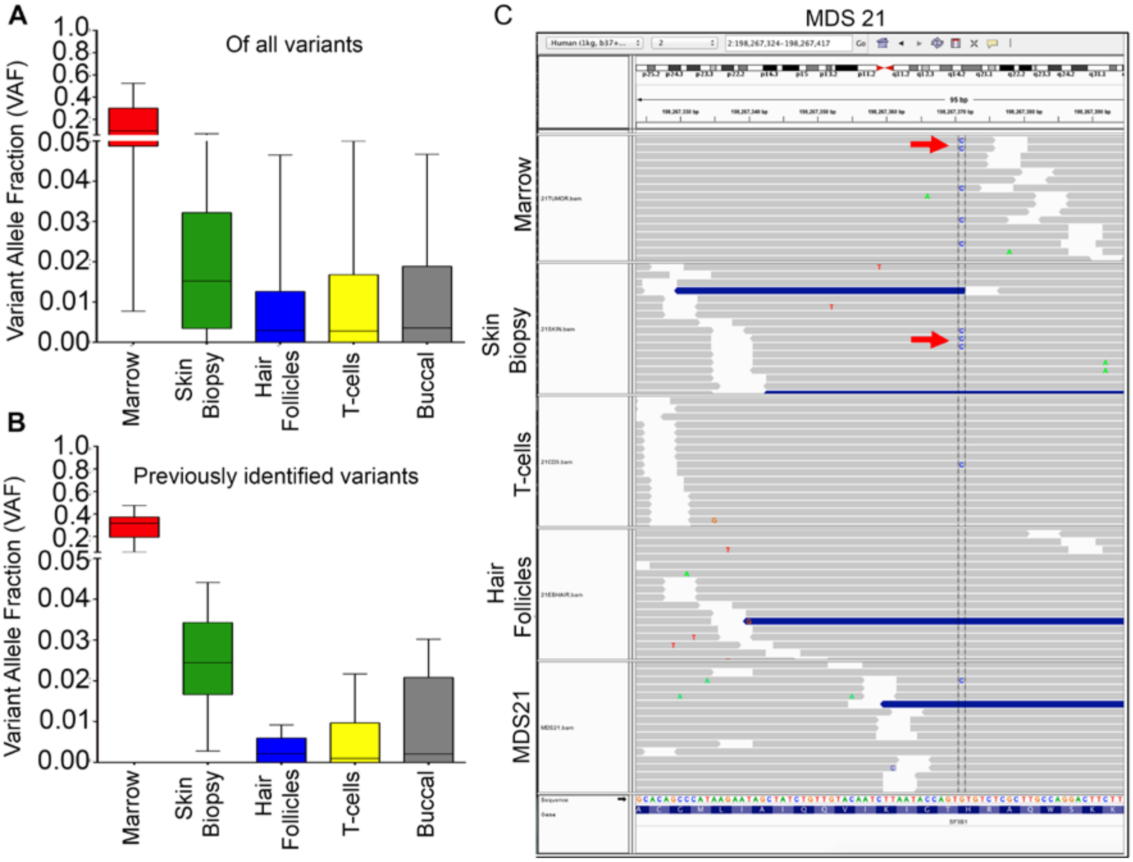


Figure S6. ***Variant allele fraction differs across tissue type.*** **(A)** Box plot of variant allele fraction of all somatic variants confirmed by targeted resequencing. This includes all variants (245 protein coding or 159 for the patients with buccal cells collected) identified in whole exome sequencing for MDS patients studied (n=16) and validated by capture-based targeted resequencing. **(B)** Box plot of variant allele fraction of somatic variants that have been previously identified in MDS (i.e., previously identified variants). These variants were confirmed by targeted resequencing and first identified in whole exome sequencing. **(C)** Representative Integrative Genomics Viewer (IGV) plots[7] of one SF3B1 mutation (H662Q, displaying sample 21). The variant AF (VAF) are: Bone marrow 24.8%, Skin: 3.7%, CD3: 0.22%, Hair follicle: 0%, Buccal (MDS21): 0.22%.

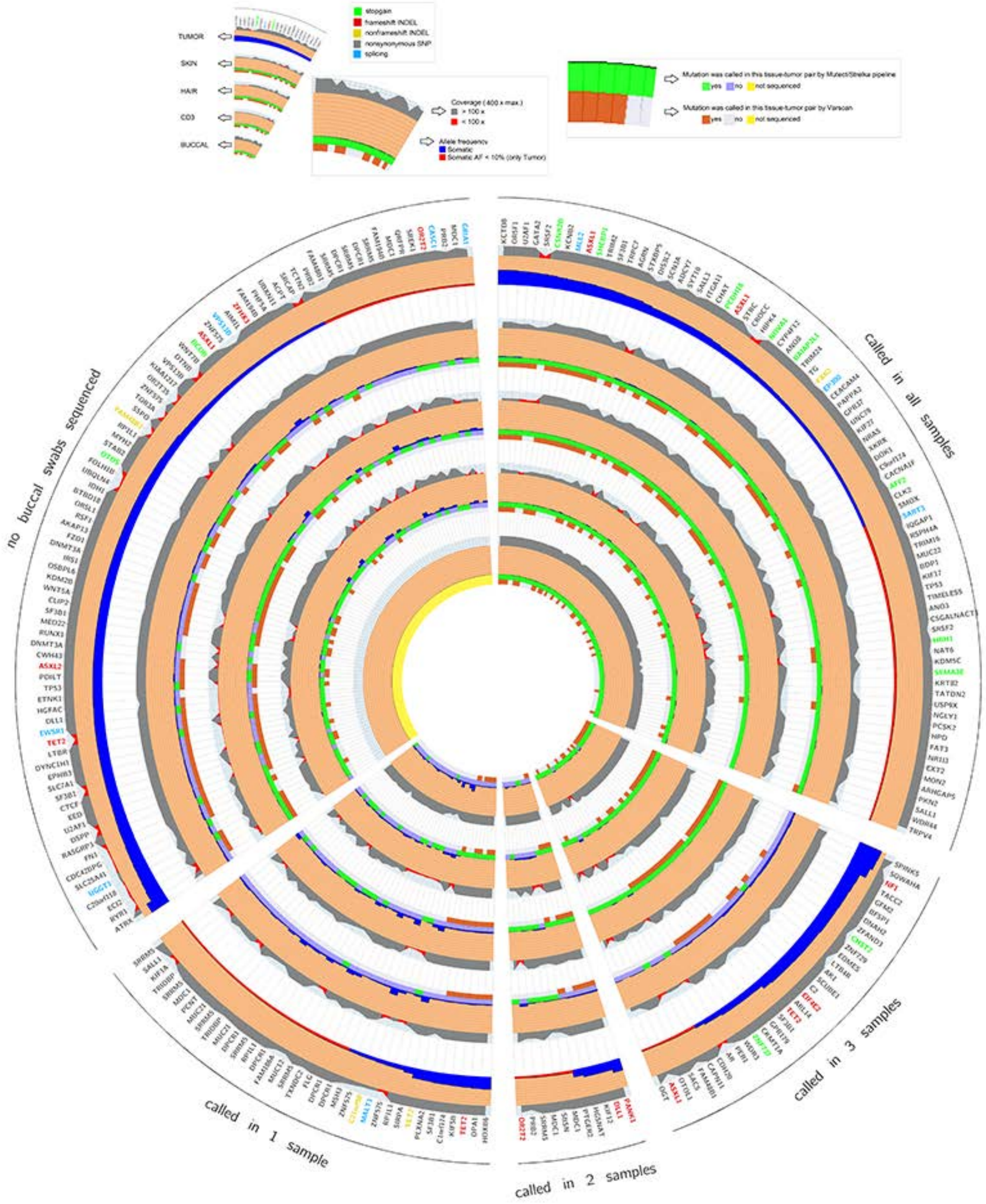


Figure S7. ***Circos diagram of tumor germline sequencing metrics.*** This Circos diagram was generated as described previously[9] and shows a representation of somatic mutations that occur in > 5% of the bone marrow. The

five wedges represent four groups of mutations with a track for VarScan calls that was used as a complementary calling method[3].

References:

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.

2. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013;31(3):213-219.

3. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 2012;28(14):1811-1817.

4. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158.

5. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-576.

6. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.

7. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.

8. Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics*. 2012;28(4):599-600.

9. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;19(9):1639-1645.