Donor cell leukemia: is reappearance of gene mutations in donor cells more than an incidental phenomenon?

Tal Shahar Gabay,[1,2] Noa Chapal-Ilani,[3] Yoni Moskovitz,[3] Tamir Biezuner,[3] Barak Oron,[3] Yardena Brilon,[3] Anna Fridman-Dror,[1] Rawan Sabah,[1,2] Ran Balicer,[4] Amos Tanay,[3,5] Netta Mendelson-Cohen,[5] Eldad J. Dann,[6] Riva Fineman,[6] Nathali Kaushansky,[3] Shlomit Yehudai-Reshef,[1#] Tsila Zuckerman[2,6#] and Liran I. Shlush[3,6#]*

[1]Hematology Research Center, Rambam Health Care Campus, Haifa; [2]Bruce Rappaport Faculty of Medicine, Technion, Haifa; [3]Department of Immunology, Weizmann Institute of Science, Rehovot; [4]Clalit Research Institute, Tel Aviv; [5]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot and [6]Department of Hematology and BMT, Rambam Health Care Campus, Haifa, Israel

[#]SY-R, TZ and LIS contributed equally as co-senior authors.

Correspondence: TSILA ZUCKERMAN - t_zuckerman@rambam.health.gov.il

doi:10.3324/haematol.2019.242347

SUPPLEMENTARY INFORMATION

Supplementary Methods

Human samples

Patients' bone marrow and peripheral blood samples were collected at diagnosis and relapse at the Rambam Health Care Campus, Haifa, Israel. Donors' samples were collected from peripheral blood prior to stem cell donation. Mononuclear cells were obtained from the bone marrow and peripheral blood using Ficoll separation and preserved in liquid nitrogen for further use. The use of human samples was approved by the Israel Ministry of Health (authorization No. 0023-15-RMB), in accordance with the Declaration of Helsinki. Patient clinical data are summarized in Supplementary Tables 1 & 2.

DNA purification

Mononuclear cells from the bone marrow aspiration or peripheral blood were thawed according to the standard protocol and DNA was extracted from both patients' and donors' mononuclear cells using Exgene^{TM} cell SV (GeneAll, cat# 106-101). DNA concentration was quantified by Qubit 3.0 fluorometer (Life Technologies,cat # Q33216). Samples that provided the DNA concentration less than 50 ng/µl underwent whole genome amplification using REPLI-g Mini Kit (QIAGEN, cat#150025). Saliva samples were collected from the patients and DNA was extracted using ORAgene-DNA purification Kit (DNA Genotek, cat# OG-500).

Chimerism analysis

Seventeen polymorphic short-tandem repeat (STR) markers were amplified using the AmpFlSTR NGM SElect Express PCR Amplification Kit (ThermoFisher, cat # 4472193). Amplified fragments were analyzed using an ABI PRISM 3500 Genetic Analyzer (Applied Biosystems) and peak areas were quantified using Gene Mapper software (Applied Biosystems). The limit of detection of the assay was approximately 1-2%; the limit of quantitation was 5%.

Cytogenetic analysis

The karyotype was prepared from bone marrow aspirates at the metaphase stage according to standard protocols. Metaphase chromosomes were G-banded using trypsin and Giemsa and karyotypes were reported according to the International System for Human Cytogenetic Nomenclature (ISCN).

Targeted Sequencing

The Molecular Inversion Probe (MIP) Capture protocol by Hiatt et al[1] was adopted and modified. Sequencing primers and MIP backbone were as described by Hiatt et.al. We planned a MIP probe set to capture ARCH defining events. We generated the MIP panel either by microarray synthesis (LCsciences) or single oligo order (Sigma Aldrich). Probes were prepared as described by Shen et al[2] and Hiatt et al, respectively. Briefly: First, the probe hybridization step was performed, followed by gap-filling and ligation steps. In order to generate a final Illumina sequencing library, the final product was mixed with barcoding PCR primers, targeting the MIP backbone, and sequenced in Novaseq6000 (Illumina) 150 paired-end run.

Variant calling

To analyze the data, a dynamic, asynchronic pipeline was constructed. It started with receipt of paired-end FASTQ files and tagging each read with its UMI barcode followed by trimming the MIP's ligation and extension arms using CUTADAPT[3]. Trimmed FASTQ files were then aligned and mapped using BWA-MEM[4] to a reference genome made out of broad's HG19 with only the relevant regions, each as a separate contig. Sam files were then sorted using Samtools sort[5] and converted to the bam format. At this stage each sample went through the process termed "family collapsing", which involved reducing every family of the same MIP that held the same UMI barcodes, resulting in new FASTQs. These FASTQs were then realigned and sorted as outlined above, Tagged using Picard's AddOrReplaceReadGroups tool[6], and processed using GATK's IndelRealigner[7]. At this point, the pipeline diverged into 2 processes for variant calling: (1) Variant calling using Varscan[8] and (2) Variant Calling using Mutect[9]. The results of the variant calling were then processed and evaluated to only include variants that appeared in two different sequencing runs of the same sample (Supplementary Table 3). Variants were further annotated using Annovar[10].

Somatic mutations were called when coverage was >1500X, Kaviar allele frequency (AF) <0.001, variant allele frequency (VAF) <0.5 and >0.005 and according to the definitions presented in Supplementary Table 4.

Clalit Database

The Clalit database includes information of patients insured by the Clalit Health Maintenance Organization (HMO) in Israel during the years 2002-2017. The Clalit dataset contains electronic medical records (EMR) of 3.45 million individuals per year on average. All data were anonymized by hashing personal identifiers and addresses followed by random sampling of data, including patient diagnoses, laboratory and medication records. This approach provided differential data analysis per patient. Diagnosis codes were acquired from both primary care and hospitalisation records, and were mapped to the ICD-9 coding system for historical reasons, with few exceptions where a partial ICD-10 coding system was used. Lab records were normalised for age and gender by subtracting raw test values from the median levels observed among all test values with matching gender and age (using a bin size of 5 years). We observed some chronological biases in lab ranges, but avoided normalising these data and instead insured patient cases and controls were matched for chronological distributions.

Defining AML cases. The EMRs were screened for all active patients (18 < age < 100) who were diagnosed with AML (ICD-9 code 205.0*) between the years 2003 and 2016.

The following exclusion criteria were applied:

1) We excluded patients with prior myeloid malignancies to omit secondary AML cases, which is consistent with the case selection for the genetic model. The following diagnoses were excluded if documented within 5 years prior to the AML diagnosis: essential thrombocythemia (ICD-9 238.71), low-grade myelodysplastic syndrome (MDS) (ICD-9 238.72) high-grade MDS lesions (ICD-9 238.73), MDS with 5q deletion (ICD-9 238.74), MDS, unspecified (ICD-9 238.75), polycythemia vera (ICD-9 238.4), myelofibrosis (ICD-9 289.83), chronic myelomonocytic leukemia (CMML) (ICD-9 206.10-206.22)

2) Patients who underwent any procedures performed on the bone marrow or spleen (ICD-10 code Z41) within 5 years prior to the first mentioning of the AML diagnosis code in their record. These patients were presumed to have an inaccurate AML diagnosis date or were misdiagnosed.

3) Patients who received medications suggestive of an alternative diagnosis, such as chronic myeloid leukaemia, lymphoid malignancy or acute promyelocytic leukaemia (APL):

- At any time prior to diagnosis: imatinib, dasatinib, anagrelide, hydroxycarbamide, asparaginase, pegaspargase, arsenic trioxide.
- At any time after diagnosis: imatinib, dasatinib, methotrexate, tretinoin, arsenic

trioxide.

- At any time after diagnosis, or more than single dose of mercaptopurine.

4) APL cases were excluded as they are not planned for allo-SCT.

5) Patients without a hospitalization record within 3 months prior to or 3 months after the diagnosis. This parameter was used as it is unlikely that an AML patient would not be hospitalized close to diagnosis. This filter reduced false positive cases and better defined the disease onset date.

We refined the estimated time of AML onset using the earliest time point when any of the following diagnoses appeared in patient's EMR: lymphoid leukemia (ICD-9 204), myeloid leukemia (ICD-9 205), leukemia of unspecified cell type (ICD-9 208).

A total of 875 AML cases in the training set were retained for further analysis. They were then validated by manual expert inspection of the complete records of 8 % of the cases.

To define the control set, we included all the individuals insured by Clalit that were not index cases. As for the present analysis, data were aggregated from a historical time window of 15 years, each control was associated with a randomised time point for evaluation. Using this approach, both index cases and controls represented a specific time point in the historical record of a patient, with matching calendric, age and gender distributions. Thus, 5,238,528 controls were used.

Out of the 875 AML patients we identified all those who survived more than two years from the initial diagnosis (N=174) and out of them all the ones who underwent allogeneic stem cell transplantation based on ICD-9 # 41.05 (N=71).

Laboratory features: Out of 2770 different types of lab tests, we selected the top 50 most frequent lab tests (Supplementary Table 5). For each lab measurement, we used median age/gender normalised test values and compared them to those of age/gender matched control population.

Supplementary Table 1: Clinical history of donors and recipients

	Patient 1	**Patient 2**
Age (years)	54	26
Sex	M	F
Primary malignancy	AML	MDS
Cytogenetics	46, XY	46, XX,+8
Initial chemotherapy	7+3	7+3
Treatment response	CR	CR
SCT 1	RIC	Myeloablative haploidentical SCT
Source	PB	PB
Stem cell manipulation	NA	TCD
Post SCT Chimerism (STR)	100%	100%
SCT 2	NA	Haploidentical SCT +Post SCT CTX
Source		PB
Chimerism (STR)	NA	100%
Donor- cell leukemia		
Time to DCL (years)	9	15
Cytogenetics	46, XX,del(6),t(6;11), (p22:q13)	47,XY,del(20)(q11.2)+(21) /46,xy,del(20)(q11.2)
Chimerism (STR)	100%	100%
	Donor 1	**Donor 2**
Age at time of donation (years)	52	73
Sex	F	M

Supplementary Table 2: CBC profile at diagnosis and relapse

	Patient 1	**Patient 2**
Diagnosis		
WBC (X10^3/µl)	2.03	2.53
Hb	10.9	7.5
MCV	103.7	96.6
RDW	17.4	21.3
PLT (X10^3/µl)	138	130
% Blast cells in BM	11	8
Relapse		
WBC (X10^3/µl)	2.22	3.73
Hb	7.1	8.4
MCV	96.2	108.6
RDW	16.9	18.7
PLT (X10^3/µl)	34	100
% Blast cells in BM	52	5
	Donor 1	**Donor 2**
WBC (X10^3/µl)	6.17	6.55
Hb	12.5	14.6
MCV	88	97.6
RDW	13.2	14.4
PLT (X10^3/µl)	173	206

Abbreviations:
AML- acute myeloid leukemia
BM- bone marrow
SCT- stem cell transplantation
CTX- cytoxan
CR- complete remission
DCL - donor cell leukemia
DNR- daunorubicin
Hb - hemoglobin
MCV - mean corpuscular volume
MDS - myelodysplastic syndrome
NA - not applicable
PB- peripheral blood
PLT- platelets
RDW- red cell distribution width
RIC- reduced intensity conditioning
SCT- stem cell transplantation
TCD- T cell depletion
WBC- white blood cells

Supplementary Table 3: Somatic mutations identified in recipient bone marrow (Dx) and donor peripheral blood pre-allo-SCT and at the time of DCL diagnosis

Patient 1

Sample	Sample date	Chromosome	Position	Mutation	Type of mutation	Location	Gene	Amino Acid Change	VAF (%)
Dx (BM)	16.7.2006	21	44524456	G>A	SNV	exonic	U2AF1	S34F	11.4
DCL (BM)	27.4.2014	21	44524456	G>T	SNV	exonic	U2AF1	S34Y	44.86
DCL (BM)	15.1.2015	21	44524456	G>T	SNV	exonic	U2AF1	S34Y	40.74
Donor (PB)	17.9.2006	2	25463238	A>T	SNV	exonic	DNMT3A	F752Y	2.96
DCL (BM)	27.4.2014	2	25463238	A>T	SNV	exonic	DNMT3A	F752Y	45.6
DCL (BM)	15.1.2015	2	25463238	A>T	SNV	exonic	DNMT3A	F752Y	43.3
DCL (BM)	15.1.2015	2	209113113	G>A	SNV	exonic	IDH1	R132C	30

Patient 2

Sample	Sample date	Chromosome	Position	Mutation	Type of mutation	Location	Gene	Amino Acid Change	VAF (%)
Dx (BM)	9.2.2000	21	44524456	G>A	SNV	exonic	U2AF1	S34F	36.86
DCL (PB)	9.2.2016	4	106158275	C>G	stopgain	exonic	TET2	S1059X	1.26
DCL (PB)	9.2.2016	20	31024747	G>A	stopgain	exonic	ASXL1	W1411X	8.26
DCL (BM)	21.8.2017	20	31024747	G>A	stopgain	exonic	ASXL1	W1411X	25.4

Supplementary Table 4: Mutation hotspots characterization

Gene name	Mutation Type	Region
ASXL1	Frameshift/nonsense/splice-site	exon 12
BCOR	Frameshift/nonsense/splice-site	whole gene
BCORL1	Frameshift/nonsense/splice-site	whole gene
BRAF	Missense	aa range p.590-615; G469
BRCC3	Frameshift/nonsense/splice-site	whole gene
CALR	Frameshift	exon 9
CBL	Missense	aa range p.345-434
CREBBP	Frameshift/nonsense/splice-site	whole gene
CSF1R	Missense	L301,Y969
DNMT3A	Frameshift/nonsense/splice-site	whole gene
DNMT3A	Missense	aa range p.292-350, p.482-614, p.634-912
EZH2	Frameshift/nonsense/splice-site	whole gene
EZH2	Missense	aa range p.617-732
GNAS	Missense	R201
GNB1	Missense	K57, I80
IDH1	Missense	R132
IDH2	Missense	R140, R172
JAK2	Missense/indel	V617F, aa range p.536-547
KDM6A	Frameshift/nonsense/splice-site	NM_021140
KRAS	Missense	G12, G13, Q61, A146
NPM1	Frameshift	exon 12
NRAS	Missense	G12, G13, Q61
PHF6	Frameshift/nonsense/splice-site	whole gene
PPM1D	Frameshift/nonsense	exon 5, 6
PRPF40B	Frameshift/nonsense/splice-site	whole gene
PTEN	Frameshift/nonsense/splice-site	whole gene
PTPN11	Missense	aa range p.58-76, p.491-510
RAD21	Frameshift/nonsense/splice-site	whole gene
SF1	Frameshift/nonsense/splice-site	whole gene
SF3A1	Frameshift/nonsense/splice-site	whole gene
SF3B1	Missense	aa range p.529-1201
SMC1A	Missense	R96, R586
SMC3	Frameshift/nonsense/splice-site	whole gene
SRSF2	Missense/deletion	P95
STAG2	Frameshift/nonsense/splice-site	whole gene
STAT3	Missense	aa range p.580-670
TET2	Frameshift/nonsense/splice-site;	whole gene
TET2	Missense	p.1104-1481, p.1843-2002
TP53	Frameshift/nonsense/splice-site;	whole gene
TP53	Missense	aa range p.95-288, P72, R337
U2AF1	Missense	S34, R156, Q157

U2AF2	Missense	aa range p.149-231, p.259-337, p.381-462
ZRSR2	Frameshift/nonsense/splice-site	whole gene
CEBPA	Frameshift/nonsense/splice-site	whole gene
RUNX1	Frameshift/nonsense/splice-site	whole gene
RUNX1	Missense	aa range p.100-440
SETBP1	missenese	868, 870
FLT3	inframe	580-820
FLT3	Missense	835, 839, 841
KIT	farmeshift	416-418
KIT	Missense	816,419
KMT2D	Frameshift/nonsense/splice-site	whole gene
NF1	Frameshift/nonsense/splice-site	whole gene
WT1	Frameshift/nonsense/splice-site	whole gene

5. Laboratory test results included in the clinical model

Parameter
Hematocrit (HCT)
Mean corpuscular volume (MCV)
Red blood cell count (RBC)
Hemoglobin (HGB)
Mean corpuscular hemoglobin (MCH)
Mean corpuscular hemoglobin concentration (MCHC)
White blood cell count (WBC)
Platelet count (PLT)
Lymphocyte percentage (LYM%)
Neutrophil percentage (NEUT%)
Eosinophil percentage (EOS %)
Monocyte percentage (MON%)
Basophil percentage (BASO %)
Absolute lymphocyte count (LYMP.abs)
Absolute neutrophil count (NEUT.abs)
Absolute eosinophil count (EOS.abs)
Absolute monocyte count (MONO.abs)
Basophiles (abs)
Mean platelet volume (MPV)
Red cell distribution width (RDW)
Creatinine - Blood
Glucose- Blood
Urea - Blood
Sodium
Potassium
Glutamic Oxaloacetic Transaminase
Glutamic Pyruvic Transaminase
MICR %
HYPO %
MACRO%
Phosphatase - Alkaline
Cholesterol
Triglycerides
LUC%
LUC
Cholesterol- HDL
Calcium - Blood
HYPER%
Uric Acid- Blood
Cholesterol- Ldl
Bilirubin Total

Albumin
Protein –Total - Blood
Phosphorus - Blood
Thyroid Stimulating Hormone (TSH)
Lactic Dehydrogenase (LDH) - Blood
Gamma Glutamyl Transpeptidase
Bilirubin - Direct
Non-HDL Cholesterol
PH-u
Specific Gravity
CK-CREAT. Kinase (CPK)
PT-INR
MICRO%/HYPO%
Vitamin B12
Iron
PT %
Prothrombine time (PT- SEC)
Chloride (Cl)
Lipemic
Icteric
Hemolytic
Hemoglobin A1C calculated
CH
Globulin
Ferritin
T4 - free
APTT-sec
folic acid
PDW
Myeloperoxidase index (MPXI)
Transferrin
PCT
Cholesterol HDL ratio
Bilirubin indirect
HCT/HGB ratio
Creatinine Urine Sample
Sedimentation Rate
Erythrocytes
Leucocytes
C-reactive protein (CRP)
RDW-CV
M.ALBUM/CREAT ratio
Amylase - Blood
MICROALBU U SAMP
Protein

Magnesium - Blood
Hemoglobin distribution width (HDW)
Fibrinogen
Sodium - Blood
Vitamin D3- 25-0H- RIA
Potassium - Blood
RDW-SD
Prostate specific antigen (PSA)
T3- free
Activated partial thromboplastin time (APTT-R)
Normoblast %
Estradiol (E-2)
Absolute normoblast count (Normoblast.abs)
Luteinizing hormone (LH)

References

1. Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*. 2013;23(5):843-854.
2. Shen P, Wang W, Chi AK, Fan Y, Davis RW, Scharfe C. Multiplex target capture with double-stranded DNA probes. *Genome Med*. 2013;5(5):50.
3. Martin M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet Journal*. 2011;17:10-12.
4. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: https://arxiv.org/pdf/1303.3997.pdf. Accessed in 09.2019; 2013.
5. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
6. Picard Tools. Available at: http://broadinstitute.github.io/picard. Accessed in 09.2019.
7. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
8. Koboldt DC, Chen K, Wylie T, et al. VarScan: variant detection in massively parallel sequencing of individ
ual and pooled samples. *Bioinformatics*. 2009;25(17):2283-2285.
9. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219.
10. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.