Supplementary Tables

Supplementary Table 1. Multivariate cox model of relapse free survival in pre-transplant MRD positive and negative patient subgroups. MRD status and chronic GVHD analysed as time-dependent variables.

Pre-Transplant MRD Status	Variable	Reference level	Hazard Ratio (95% CI)	P
MRD positive	*FLT3* status - Present	Absent	1.23 (0.72, 2.12)	0.45
	Cytogenetic risk group - Adverse	Favourable/Intermediate risk	1.14 (0.67, 1.97)	0.63
	Post-transplant MRD status (time dependent) - Positive	Negative	7.03 (3.82, 12.9)	<0.001
	Chronic GvHD (time dependent) - Yes	No	5.85 (2.41, 14.2)	<0.001
MRD negative	*FLT3* status - Present	Absent	1.29 (0.78, 2.12)	0.32
	Cytogenetic risk group - Adverse	Favourable/Intermediate risk	1.54 (0.94, 2.51)	0.086
	Post-transplant MRD status (time dependent) - Positive	Negative	7.01 (4.09, 12.0)	<0.001
	Chronic GvHD (time dependent) - Yes	No	6.65 (2.94, 15.0)	<0.001

Supplementary Table 2
Comparison of the characteristics between patients with mixed (<95%) versus full (≥95%) donor T cell chimerism at 3 months post-transplant.

		T cell chimerism at 3 months				
		Mixed N (%)	**Full N (%)**	**Missing throughout* N (%)**	**Overall N (%)**	**P**
Treatment arm	FLAMSA-BU	17 (33)	52 (50)	21 (55)	108 (50)	0.078
	Fludarabine/busulphan/ATG	16 (31)	36 (35)	9 (24)	63 (29)	
	Fludarabine/melphalan/ alemtuzumab	10 (19)	11 (11)	6 (16)	30 (14)	
	Fludarabine/busulphan/ alemtuzumab	9 (17)	4 (4)	2 (5)	15 (7)	
Age	<=60 years	29 (56)	60 (58)	25 (66)	126 (58)	0.80
	>60 years	23 (44)	43 (42)	13 (34)	90 (42)	
Sex	Female	19 (37)	49 (48)	13 (34)	91 (42)	0.41
	Male	33 (63)	54 (52)	25 (66)	125 (58)	
Underlying disease	AML	35 (67)	69 (67)	22 (58)	144 (67)	0.74
	MDS	17 (33)	34 (33)	16 (42)	72 (33)	
Cytogenetic risk - AML patients	Adverse risk	11 (31)	18 (26)	9 (41)	44 (31)	0.83
	Intermediate Risk	23 (66)	44 (64)	13 (59)	92 (64)	
	Favourable Risk	1 (3)	6 (9)		7 (5)	
	Unknown		1 (1)		1 (1)	
Disease status (AML only)	CR1/CR2	33 (94)	67 (97)	20 (91)	138 (96)	0.53
	Primary refractory	2 (6)	2 (3)	2 (9)	6 (4)	
FLT3	Absent	21 (40)	41 (40)	16 (42)	87 (40)	0.94
	Present	8 (15)	20 (19)	4 (11)	37 (17)	
	Unknown	23 (44)	42 (41)	18 (47)	92 (43)	
NPM1	Absent	20 (38)	43 (42)	16 (42)	88 (41)	0.97

		T cell chimerism at 3 months				
		Mixed N (%)	**Full N (%)**	**Missing throughout* N (%)**	**Overall N (%)**	**P**
	Present	8 (15)	18 (17)	4 (11)	35 (16)	
	Unknown	24 (46)	42 (41)	18 (47)	93 (43)	
IPSS (MDS only)	Standard risk (<=2)	14 (100)	30 (97)	14 (100)	60 (97)	1
	High risk (>2)		1 (3)		2 (3)	
Donor type	Sibling	16 (31)	17 (17)	7 (18)	45 (21)	0.22
	Unrelated	36 (69)	86 (83)	31 (82)	171 (79)	
Stem cell source	Peripheral blood	51 (98)	100 (97)	34 (89)	208 (96)	0.21
	Bone marrow	1 (2)	3 (3)	4 (11)	8 (4)	
Pre transplant MRD	Positive	16 (31)	17 (17)	9 (24)	43 (20)	0.42
	Negative	23 (44)	55 (53)	22 (58)	113 (52)	
	Missing	13 (25)	31 (30)	7 (18)	60 (28)	
DLI	Number receiving DLI (prior to relapse)	20 (38)	4 (4)	6 (16)	30 (14)	<0.0001
GVHD	Acute GVHD Grade 2-4	21 (40)	43 (42)	10 (26)	82 (38)	0.40
	Chronic GVHD	21 (40)	38 (37)	8 (21)	67 (31)	0.18

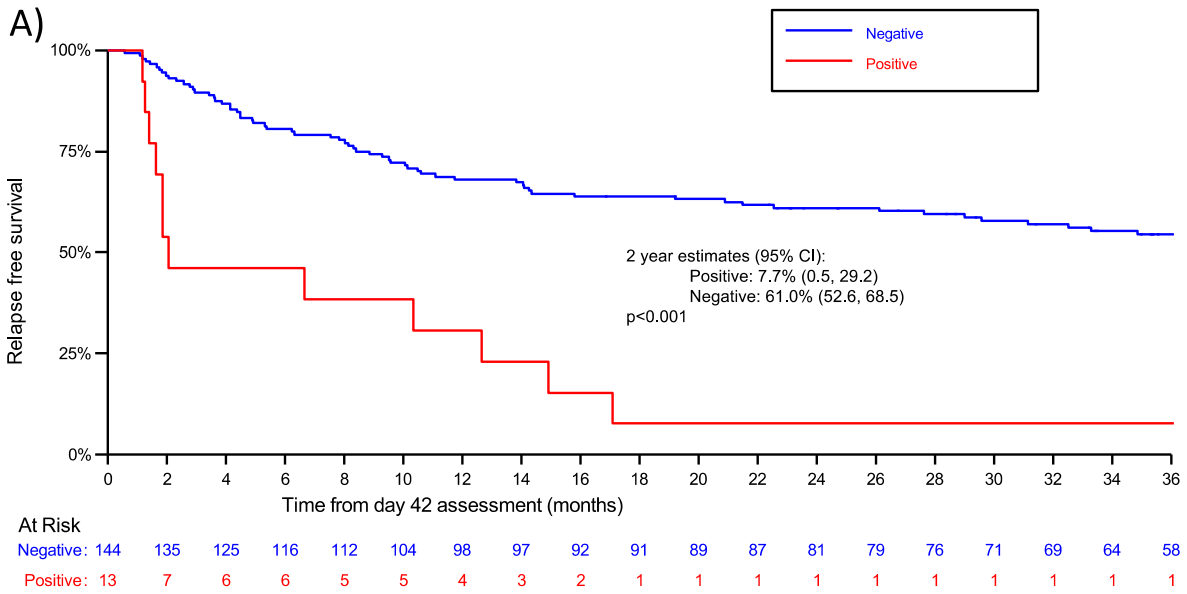
Legend. *'Missing throughout' includes the 12 of 216 transplanted patients who died or relapsed up to day+42

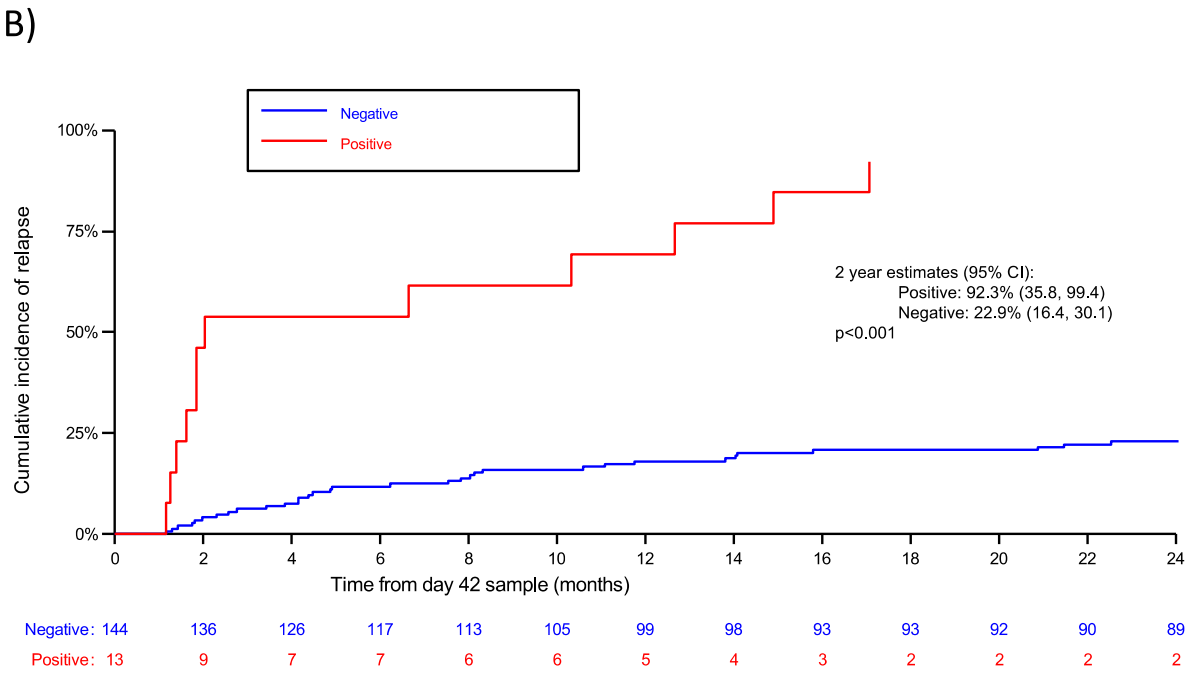
Abbreviations. DLI, donor lymphocyte infusion; GVHD, graft versus host diseas

Supplementary Figures

Supplementary Figure 1

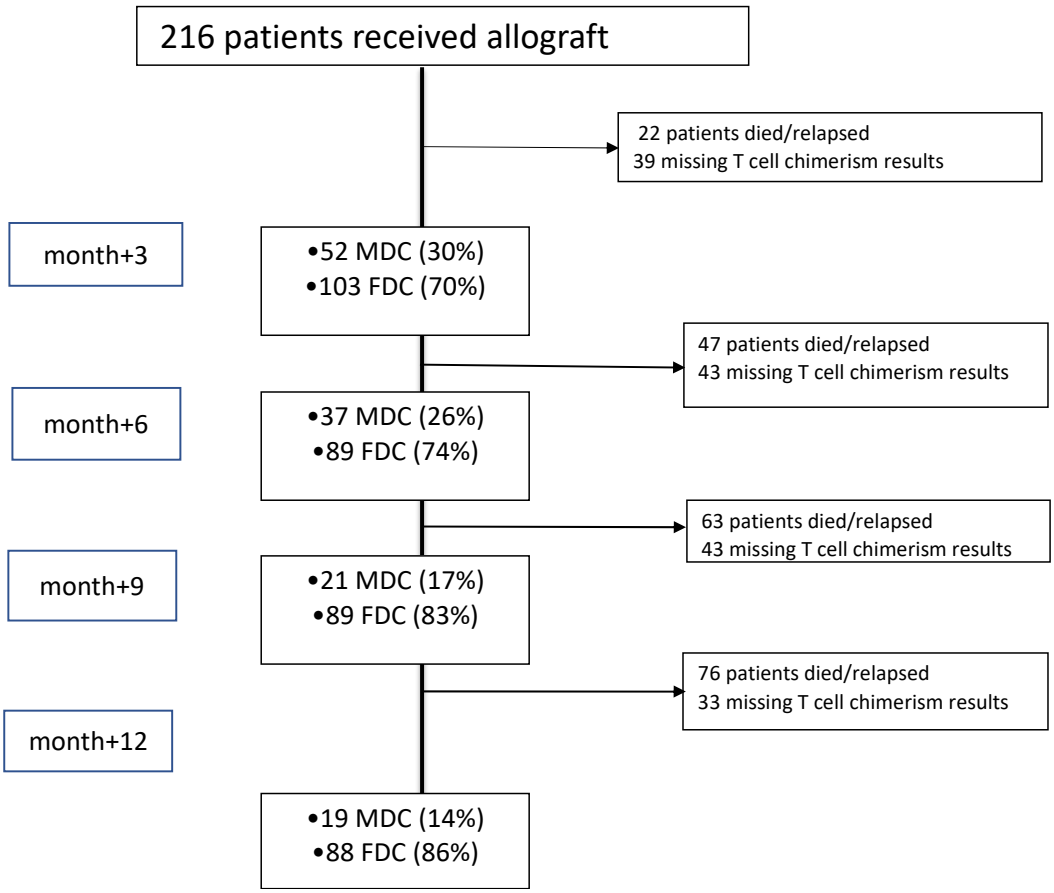
A) RFS and B) CIR in patients who were MRD positive compared to negative from D+42 post transplant MRD assessment.



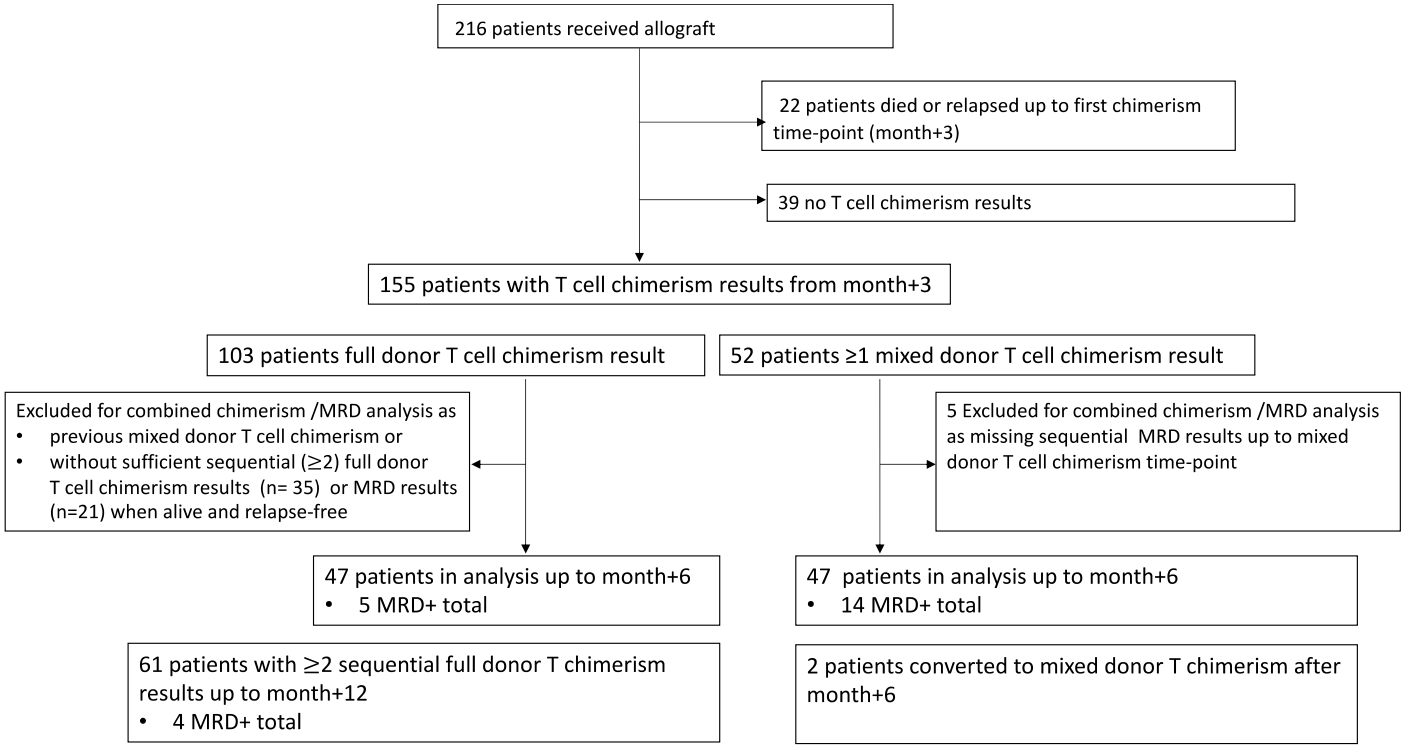


Supplementary Figure 2. Flow charts for:
A. T cell chimerism analysis B. Combined T cell chimerism with MRD analysis
Abbreviations: MRD, measurable (minimal) residual disease;

A. T cell chimerism



B. Combined T cell chimerism with MRD



Supplementary Methods

Use of Prophylactic Donor-Lymphocyte Infusion

Donor-lymphocyte infusion (DLI) was administered as per trial protocol.

For the control arm:
DLI was administered at 6 months post-transplant in patients with mixed T cell chimerism (defined as <98% donor T cells observed in 2 consecutive results) who have discontinued immunosuppression and have no evidence of active GvHD. DLI was scheduled every 3 months using an escalating dose regimen until 100% donor T cell chimerism is achieved. The following DLI dosing schedule was used:-

Time point after transplant (months)	CD3+ cells/kg
6	1×10^6
9	3×10^6
12	1×10^7
15	3×10^7
18	1×10^8

DLI for experimental arm:
DLI was administered at 4 months post-transplant in patients in remission with mixed T cell chimerism (defined as <98% donor T cells observed in 2 consecutive results), if there were no history of GvHD and have discontinued immunosuppression. DLI will be administered every 2 months using an escalating dose regimen until 100% donor T cell chimerism is achieved. The following DLI dosing schedule should be used:

Time point after transplant (months)	CD3+ cells/kg
4	1×10^6
6	5×10^6
8	1×10^7

Lineage specific chimerism analysis

Lineage specific chimerism analysis was performed by local laboratories according to a standardised, nationally accredited quality assurance system.[1] For example, CD3 + cells were separated from whole peripheral blood using MACS (Miltenyi Biotec) and DNA

extracted from cell suspensions using the EZ1 Advanced automated extractor (Qiagen). Proportions of donor/recipient chimerism were determined using multiplex PCR of a bespoke panel of 16 fluorescently labelled, highly informative microsatellite markers, using 25-50ng DNA template. PCR cycles were limited to the exponential phase (25 cycles) to ensure accurate quantification. PCR product was diluted 1:2 and 1uL loaded onto an AB3500xL Genetic Analyzer (ThermoFisher) with 12 second injection time. ChimerMarker (SoftGenetics) automated analysis software was used to calculate relative levels of donor and recipient in each sample based on peak heights within informative markers, with a sensitivity of 1-2% for the minor population.

Statistical methods

Cox proportional hazard model has been employed for the multivariate regression analysis. When carried out assessment of proportional hazards for each covariate was assessed visually using Kaplan Meier curves and if crossing curves were observed then log(-log) plots were used to assess the assumption formally. Assumptions held for all models presented in the analysis. No variable selection was employed in the regression analysis. No imputation for missing data was carried out and therefore only patients with full datasets were included in the model analysis. Cytogenetic risk scores were classified as adverse for AML[2] and MDS[3] based on validated scoring systems.

Multiparameter Flow Cytometry (MFC) detection of MRD

Patients' samples were sent by overnight mail to the reference laboratory. Following ammonium chloride lysis, bone marrow nucleated cells were labelled with antibody panel shown in Supplementary Methods Table 1 for flow cytometric MRD analysis as previously described[4,5]. Where performed, tube 3 provided information on myeloid maturation profile and an estimate of hemodilution (by CD11b/CD13 profile) but was not included in MRD analysis. Tube 4 was applied to detect leukemic stem cell (LSC) immunophenotypic aberrancies (from CLL1/ CD45RA/CD123 expression). Cell acquisition was performed on a FACSCanto (BD Biosciences) flow cytometer. Acquisition was set for 500,000 to 1 million cells or as many cell events as possible for MRD samples. Post-acquisition analysis of the flow cytometry data was performed (blinded to clinical data) using FlowJo software (Treestar Inc). Data review for analyses included periodically updated reference control bone marrow profiles. Viability, acquisition and autofluorescence artefact and

hemodilution were assessed in acquisition generated flow cytometry standard (FCS) data files. Assay limit of detection was 0.05% of leukocytes.

MRD Analysis

Computational unsupervised analysis was applied to the flow cytometrically generated immunophenotypic data to exclude any potential variation in MRD results from subjective interpretation. This analysis approach was previously demonstrated to be at least equivalent for relapse prediction compared to manual flow cytometric MRD analysis[4]. Blasts were categorised as CD34+ and /or CD117+ cells to minimise variability that might be introduced by more 'inclusive' blast gating. Blast FCS files for computational clustering packages were generated using FlowJo v.10 software (FlowJo LLC, Ashland, OR, USA). FCS files were gated by time (to identify and filter acquisition artefacts) then cleaned by exclusion of debris and dead cells exclusion. Blast events were selected by sequential gating (leukocytes by CD45+, mononuclear, CD34+ and/or CD117+, intermediate CD45). At least 1000 blast events were required to proceed to unsupervised analysis. Random down-sampling of gated blasts to 2700 and/or 1000 (if <2700 blasts) events was applied. Blast numbers were ≥ 2700 in 81% of patient samples. Samples with <1000 blast events (6.6% of samples) were defined as MRD negative if adequate and MRD negative by standard analysis, or otherwise as non-assessable. All blast files including those from control samples in any unsupervised analysis run had a consistent number of blasts (2700 or 1000) to remove any dominating effect from a large over-represented population.

Clustering algorithms and dimensionality reduction for initial visualisation were performed using the open-source R package, Cytofkit v1.12.0 (github.com/ JinmiaoChenLab/cytofkit)[6]. Blast FCS files were uploaded to Cytofkit via GUI interface in R-studio (R version 3.5.2), and parameters of interest were selected before analysis was run. Computational clustering algorithms take into account all selected markers simultaneously rather than biaxial plots and group cells into subpopulations (clusters) based on similarity of marker expression. As multiple sample files can be batched and simultaneously analysed using the Cytofkit package, each batch included a set number of controls per tube in addition to the blast files from AML patient samples. By providing a reference range from a set of control blast populations for each analysis run, this allowed a 'different-from normal' computational analysis for detection of aberrant leukemic subpopulations. Phenograph clustering[7,8] and t-distributed stochastic neighbour embedding (tSNE) dimensionality reduction was performed with the following settings;

merge method ‘all’, transformation method autoLgcl, Seed 42, tSNE Perplexity 30, Max iterations 1000, RPhenograph k value 80. Cytofkit run-time for Phenograph clustering was approximately 40 minutes for 2700 event analysis of a group of 40-50 control samples, and 50 test samples (per antibody combination). Each run defined between 15 and 28 distinct clusters (median 21). Each computational analysis run generated a modified FCS file (including tSNE parameters and cluster ID tags) per control/patient sample and CSV files stating cluster frequencies. Modified FCS files were visualised in FlowJo, and converted to CSV format for integration into the statistics pipeline that applied criteria defining different from normal immunophenotypic aberrancies. Contaminating auto-fluorescent populations and B-cell progenitors were excluded from MRD quantification. The event threshold for quantifiable populations / clusters was set at ≥34 blast events (applying Poisson statistics for a CV of 20% and incorporating 95% CI to account for measurement error[9,10]) and 10% of the cluster.

Blast cells from MRD samples were defined as aberrant when part of a Phenograph cluster that was ≥5% of patient blasts and not quantifiable in any of the control samples (**criteria 1**) or, if in a cluster also containing control cells, when blasts had significantly different marker expression of ≥1 marker compared to the cluster control cell distribution (**criteria 2**).

Aberrancy by criteria 2 was determined using statistically defined gating thresholds for each marker[11]. Thresholds were defined using the $10^{th}/90^{th}$ percentiles of marker expression (fluorescent intensity) of the concatenated control cells within a cluster. Flow cytometry data is susceptible to variable electronic noise at the lower end of the fluorescence scale, and clustering algorithms may define artificially tight populations from e.g. compensation artifacts. To account for this, data were scaled to remove negative values, and a multiplier/divisor (modifier) proportional to the inverse of the trimmed cluster standard deviation ($1/SD_{TRIM}$) was applied to calculate a threshold with increased stringency. To account for outlier control phenotypes, the calculated gating threshold for a cluster was applied to all control blast files.

Cluster MRD positivity by DfN was reported when only test cells were quantifiable as aberrant by calculated $10^{th}/90^{th}$ percentiles of marker expression of the concatenated control cells (***criteria 2a****,* example Supp Methods Figure 1.a). If control cells were also quantifiable, a more stringent threshold was applied defined from the 10^{th} or 90^{th} percentile of fluorescent intensity for the control with the highest number of quantifiable

events above the original threshold. Cluster MRD positivity was reported when test cells were quantifiable as aberrant by this adjusted $10^{th}/90^{th}$ percentiles of marker expression (***criteria 2b***, example Supp Methods Figure 1.b-c).
Over-expression of either CD56 or CD7 in ≥5% of blasts by this pipeline was identified as a highly specific aberrancy (absent in all 50 controls).

Frequency of blasts fulfilling the above DfN criteria by unsupervised analysis was summated as % leukocytes for each MRD tube (tubes 1/2/stem cell). MRD result was reported as positive when MRD was above the limit of quantitation (defined as LOB+(1.625*SD)) and in ≥5% of blasts in at least 2 of the 3 tubes or for CD56 /CD7 aberrancy (from tube 2). Most positive results were ≥0.1% MRD by manual analysis.

Detection of HLADR negative aberrant blast populations in MRD positive samples
Quantifiable aberrant (DfN) clusters detected by markers that included HLADR (tube 1) were screened for HLADR negativity in FlowJo. HLA-DR negativity was defined by a set gate at fluorescence intensity 500 and reported as present (MRD positive /HLADR negative) if quantifiable by event thresholds (≥34 events, ≥10% cluster).

References

1. Clark JR, Scott SD, Jack AL, et al. Monitoring of chimerism following allogeneic haematopoietic stem cell transplantation (HSCT): Technical recommendations for the use of Short Tandem Repeat (STR) based techniques, on behalf of the United Kingdom National External Quality Assessment Service for Leucocyte Immunophenotyping Chimerism Working Group. *British Journal of Haematology*. 2015;168(1):26-37.
2. Grimwade D, Hills RK, Moorman AV, et al. Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. *Blood*. 2010;116(3):354-365.
3. Greenberg PL, Tuechler H, Schanz J, et al. Revised International Prognostic Scoring System for Myelodysplastic Syndromes. *Blood*. 2012;120(12):2454-2465.
4. Craddock C, Jackson A, Loke J, et al. Augmented Reduced-Intensity Regimen Does Not Improve Postallogeneic Transplant Outcomes in Acute Myeloid Leukemia. *J Clin Oncol*. 2021;39(7):768-778.
5. Freeman SD, Hills RK, Virgo P, et al. Measurable Residual Disease at Induction Redefines Partial Response in Acute Myeloid Leukemia and Stratifies Outcomes in Patients at Standard Risk Without NPM1 Mutations. *J Clin Oncol*. 2018;36(15):1486-1497.

6. Chen H, Lau MC, Wong MT, Newell EW, Poidinger M, Chen J. Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. *PLoS Comput Biol*. 2016;12(9):e1005112.
7. Levine JH, Simonds EF, Bendall SC, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. 2015;162(1):184-197.
8. van der Maaten L, Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*. 2008;9(86).
9. Hedley BD, Keeney M. Technical issues: flow cytometry and rare event analysis. *Int J Lab Hematol*. 2013;35(3):344-350.
10. Altman DG. Practical Statistics for Medical Research: Chapman & Hall/CRC; 2006.
11. McCarthy N, Loke J, Andrew G, et al. Validation and clinical applicaton of an unsupervised analysis approach to measurable residual disease testing in acute myeloid leukemia EHA2021. Vol. 5:S2: HemaSphere; 2021.

Supplementary Methods Table 1. Flow cytometric MRD Antibody Panel

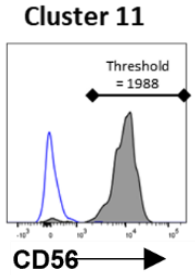
Tube No.	FITC	PE	PerCP	PECy7	APC	APC H7	Horizon V450	Horizon V500	Brilliant Violet V421
1	**HLADR** L243 (BD)	**CD13** L138 (BD)	**CD34** 8G12 (BD)	**CD117** 1042D2 (BD)	**CD33** P67.6 (BD)	**CD45** 2D1 (BD)	**CD19** SJ25C1 (BD)		
2	**CD38** HB7 (BD)	**CD56** MY31 (BD)	**CD34**	**CD117**	**CD33**	**CD45**	**CD7** M-T701 (BD)		
3	**CD13** WM-47 (Dako, Alere)	**CD11b** ICRF44 (BD Pharmingen)	**HLADR** L243 (BD)	**CD117**	**CD14** MoP9 (BD)	**CD45**			
4	**CLL1** (CLEC12A, BD)	**CD123** 7G3(BD)	**CD34**	**CD117**	**CD19** SJ25C1 (BD)	**CD45RA**		**CD45**	**CD38**

BD – Becton Dickinson Biosciences, Oxford, United Kingdom
BD Pharmingen – Becton Dickinson Biosciences - Pharmingen, Oxford, United Kingdom
Dako - Dako Agilent Technology

Supplementary Methods Figure 1. Examples of detection criteria for Unsupervised Different from Normal analysis

a.

Criteria 2a

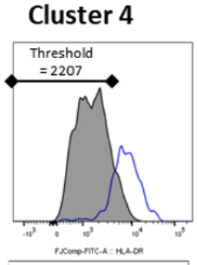


Cluster control cells - Total
Cluster test cells

Cluster 11 56+
= 60% blasts
No control cells with 34 events above calculated threshold

b.

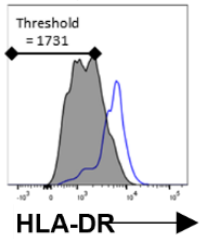
Criteria 2a



Cluster control cells - Total
Cluster test cells

Cluster 4 HLADRwk
= 33.9% blasts

Criteria 2b

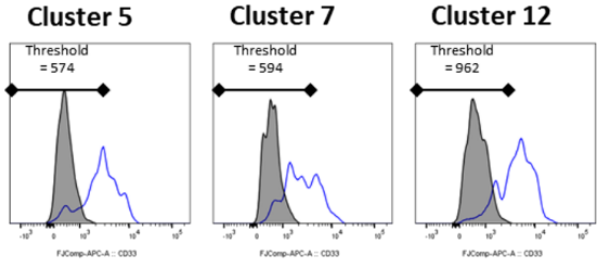


Cluster control cells – Lowest control only
Cluster test cells

Cluster 4 remains HLADRwk
= 25.8% blasts

c.

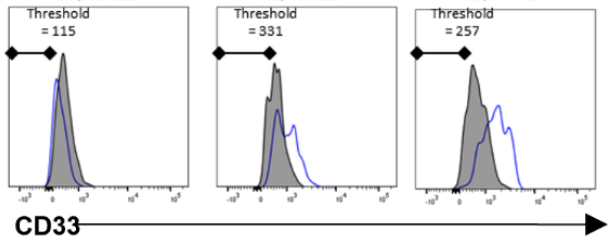
Criteria 2a



Cluster control cells - Total
Cluster test cells

Clusters 5/7/12 CD33wk
= 34% blasts

Criteria 2b



Cluster control cells – Lowest control only
Cluster test cells

Cluster 7 minor CD33wk
= 2.7% blasts

Supplementary Methods. Figure. 1: The calculated thresholds for marker under-/over-expression are represented as gates in FlowJo software. Cluster test control cells are depicted by grey and blue outline respectively.

Examples **a.** Cluster MRD positive for test cells as no individual control files had quantifiable populations over-expressing CD56 above the upper threshold in cluster (cluster 11 of sample run) (*criteria 2a*);

b. Cluster test cells potential MRD positive by HLA-DR under-expression (*criteria 2a*) but control cells detectable below the lower threshold for HLA-DR expression in cluster, so threshold revised to 10^{th} percentile of lowest control (*criteria 2b*). Cluster remains MRD positive for test cells by HLA-DR under-expression at 25.8% of test blasts;

c. In clusters 5, 7 and 12 of sample run, although test cells are potentially MRD positive by CD33 under-expression (*criteria 2a*), control cells are detectable below the lower threshold for CD33 expression so threshold revised to 10^{th} percentile of lowest control (*criteria 2b*). Clusters 5 and 12 are negative by adjusted threshold, while cluster 7 remains positive for CD33 under-expression in 2.7% of test blasts.