**Supplementary information**

**Mutation signatures of pediatric acute myeloid leukemia and normal blood progenitors associated with differential patient outcomes**

Arianne M. Brandsma1, Eline J.M. Bertrums1, Markus J. van Roosmalen1, Damon A. Hofman1,Rurika Oka1, Mark Verheul1, Freek Manders1, Joske Ubels1, Mirjam E. Belderbos1, Ruben van Boxtel1

Affiliations:

1. Princess Máxima Center for Pediatric Oncology and Oncode Institute, Heidelberglaan 25, 3584CS Utrecht, The Netherlands

Corresponding author:

Ruben van Boxtel, PhD

R.vanBoxtel@prinsesmaximacentrum.nl

**Supplementary Information includes:**

Supplementary Figures S1-S6

Figures Legends for Supplementary Figures S1-S6

Supplementary Tables S1-S10

**Supplementary Figure Legends**

**Supplementary Figure S1. VAF plots and whole genome sequencing metrics**

(A) Statistical modeling of the distribution of clonal and sub-clonal mutations by a Bayesian Dirichlet process of two HSPC clones. The empiric histogram of mutations is shown in grey, with the fitted distribution as a black line. The 95% posterior confidence intervals for the fitted distribution are shown in pale pink and dashed red line indicates a variant allele frequency (VAF) cutoff of 0.3. plot for one example HSPC. (B) Schematic representation of the HSPC clonal culture and AML a simplified model of clonal evolution from the most common recent ancestor (MRCA) resulting in whole genome sequencing (WGS) data and VAF plots. (C,D) Violin plots of the total genome coverage (C) or genome coverage ≥5 (minimum of 5 reads per position) (D) of all WGS data in this manuscript comparing samples sequenced at 15X or 30X genome coverage. As expected, the p value indicates a significant difference between 15X and 30X samples (two-sided t test). (E,F) Comparison of the number of base substitutions (E) or indels (F) per genome of HSPCs sequenced at 15X or 30X genome coverage. p value indicates a non-significant difference between 15X and 30X samples (two-sided t test). (G) Boxplot depicting the number of base substitutions per genome for HSPCs of PMC16332 comparing 15X (n=7) with 30X (n=3) genome coverage clones. In addition, the 30X HSPC clones were randomly down sampled (DS) to 15X genome coverage (only half of the reads were used) and the number of identified base substitutions depicted. p value indicates non-significant differences between 15X, 15X-DS and 30X samples (Wilcoxon Mann-Whitney test).

**Supplementary Figure S2. Flow cytometry sorting strategy and neutrophil recovery after chemotherapy**

(A) Representative FACS plot for purifying AML blasts and HSC and MPP from pAML mononuclear bone marrow cells at diagnosis. Singlet Lin-CD34+CD38-CD45RA- (MPP) or Lin-CD34+CD38-CD45RA-CD90+ (HSC) were identified, as well as Lin-CD33+CD38+ AML blasts for PMC22813. (B) Statistical modeling of the distribution of clonal and sub-clonal mutations by a Bayesian Dirichlet process of example HSPC clones and the corresponding AML blasts. The empiric histogram of mutations is shown in grey, with the fitted distribution as a black line. The 95% posterior confidence intervals for the fitted distribution are shown in pale pink and dashed red line indicates a variant allele frequency (VAF) cutoff of 0.3 for clonal mutations. (C) Outgrowth of single cell HSPC clones sorted from 6 pAML diagnosis bone marrow. t(8;21) (n=3) and ‘rest’ (n=3) AML (1 cytogenetic normal, 1 MLL-rearranged, 1 NUP98-NSD1) samples were analyzed. p value indicates a statistical difference between the clonal outgrowth of HSPCs from t(8;21) pAML and rest (two-sided t test). (D) Overview of treatment protocols for cohort of 273 pAML cases diagnoses in the Netherlands between 2005 and 2019. (E) Risk group stratification and AML subtype per age group is depicted for pAML cases in (D). (G,H) Neutrophil recovery time in days after start first course (G) or second course (H) of chemotherapy, measured as the interval between start of chemotherapy and absolute neutrophil count >0.5 × 109/L. CN, cytogenetic normal, ­­inv(16), inversion in chromosome 16, other, AML subtypes not categorized by any of the other groups.

**Supplementary Figure S3. MIP analysis of PMC22813 and PCR analysis of PMC21636**

(A) Correlation between variant allele frequency (VAF) of whole genome sequencing (WGS) and molecular inversion probes (MIP) analyses performed on PMC22813AML DNA on 87 positions. Dashed lines indicate VAF cutoff of 0.1. (B) PCR analysis for MLL-fusion specific product (F) or wild-type MLL specific product (WT) in DNA of AML blasts or n=15 HSPC clones of PMC21636. Data are representative of n=92 HSPCs analyzed in total. Only the AML blasts show the fusion-specific band at 201 bp, while all samples show the WT band, indicating that none of the HSPCs analyzed share the MLL-fusion driver event with the AML blasts. Numbers on the left indicate the size of the marker bands in base pairs.

**Supplementary Figure S4. Mutations in pAML**

(A,B) Ratio between the observed and expected number of base substitutions (A) or indels (B) per genome. Expected number is extrapolated from the linear mixed model in Fig. 1B,C. p value indicates a statistical difference between AML and patient-matched normal HSPCs (HSPC) (Mann-Whitney test). (C) 96-trinucleotide spectrum for all mutations in AML of each PMC pAML patient.

**Supplementary Figure S5. Somatic mutations in pAML and survival analysis**

(A) Comparison between number of AML drivers in “above” baseline and “on” baseline’ pAML of TARGET-21 and PMC patients combined. p value indicates a significant difference between “above” and “on” baseline patients (Mann-Whitney test). (B,C) Kaplan-Meier survival curves for 5-year event-free (B) and overall (C) survival of pAML patients classified as “above” or “on” healthy baseline in the TARGET-21 and PMC cohorts combined. p value indicates a trend towards significance (B) or no significant difference (C) between “above” and “on” baseline patients (Log-rank Mantel-Cox test). (D) Correlation of the number of base substitutions accumulated per genome with age of the assessed patients. Data from 103 pAML patients of the TARGET-20 cohort are depicted. Each data point represents a single bulk AML sample. Dotted line indicates the mutation accumulation in healthy individuals, defined by the linear mixed model in Fig. 1B. The 95% probability interval of the linear mixed model is depicted in gray. Samples above this 95% probability interval are classified as “above” pAML.

**Supplementary Figure S6. Survival analyses of TARGET-20 and gene expression analyses of TARGET-21 pAML patients**

(A) pAML subtypes of 103 TARGET-20 patients classified as “above” (n=25) or “on” (n=78) healthy baseline. (B) Multivariable models for event-free and overall survival of above and on healthy baseline pAML. Hazard ratios and p values are derived from the Cox proportional hazard model using 103 TARGET-20 pAML patients. Gray symbols indicate reference levels. \*: p < 0.05; \*\*: p < 0.01. (C,D) Kaplan-Meier survival curves for event-free (C) and overall (D) survival in 87 TARGET-20 pAML patients classified as “above” or “on” healthy baseline, excluding t(8;21) pAML. p value indicates a significant difference between “above” and “on” baseline patients (Log-rank Mantel-Cox test). (E) Confusion matrix of the random forest model trained on healthy HSC and MPP somatic mutation data. Class error depicts the out-of-bag error of the model. (F). Linear regression model was trained on 75% of the TARGET pAML data, and its performance assessed using the last 25% of the TARGET and 6 PMC pAML patients. Pearson correlation indicates a significant correlation between the true and predicted ratio to healthy baseline. Selected features of the model are depicted in the inset. (G) Differential gene expression of 15 TARGET-21 pAML patient classified as “above” or “on” healthy baseline, where “above” was compared to “on” as the reference. Blue indicates significantly up- (log2 fold change > 0.585) or downregulated (log2 fold change < -0.585) genes (padj <0.05). (H) Mean normalized expression of 19 *HOX* genes, depicted in (I), for each pAML sample of the TARGET-21 cohort. p value indicates a significant difference between “above” and “on” baseline patients (Mann-Whitney test). (I) Heatmap depicting the expression of 19 *HOX* genes of TARGET-21 pAML patients, expression mean is normalized to 0. (J) Gene set enrichment plots of the indicated gene sets using log fold change shrinkage data. NES indicates the normalized enrichment score, which corrects for multiple testing. A negative NES indicates genes enriched in “on” baseline pAML of TARGET-21, as these were used as reference. (K) Normalized expression of MEIS1 in of TARGET-20 pAML “above” or “on” healthy baseline. p value indicates a significant difference between “above” and “on” baseline patients (Mann-Whitney test).