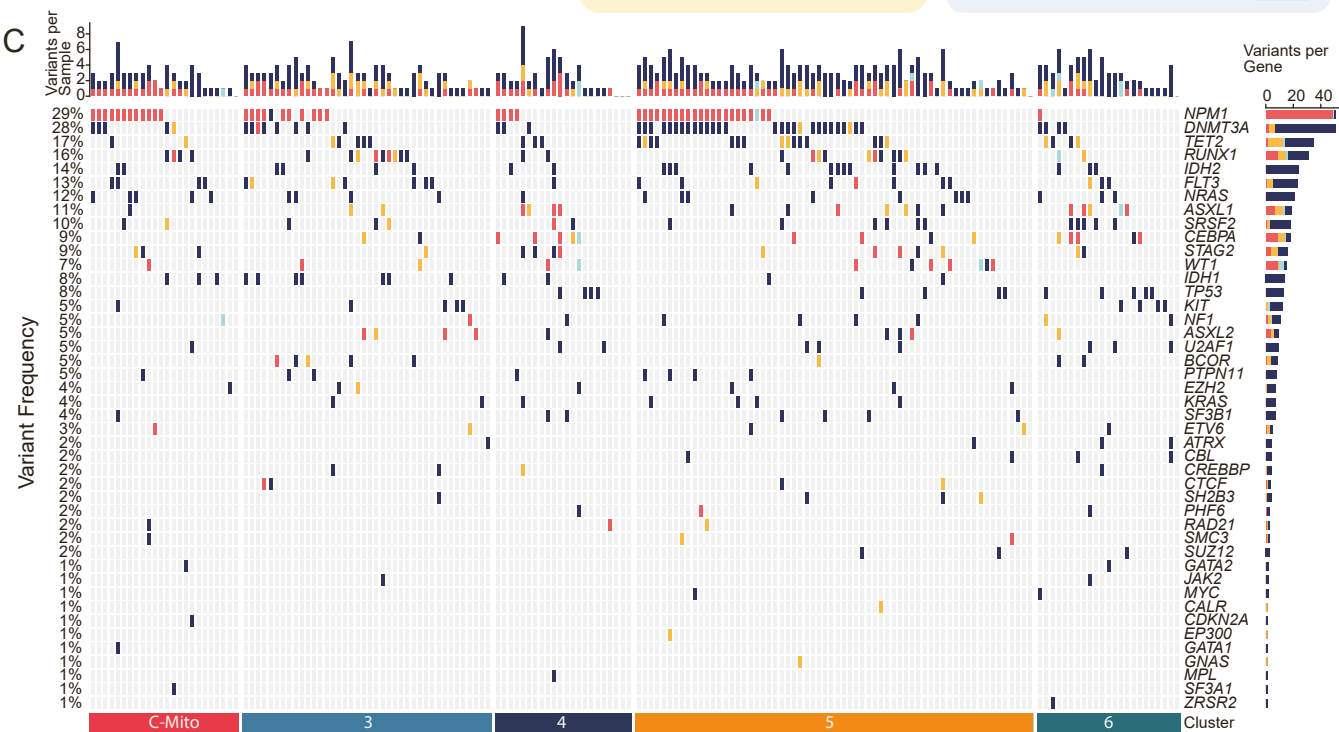


Supplemental information

The proteogenomic subtypes of acute myeloid leukemia

Ashok Kumar Jayavelu, Sebastian Wolf, Florian Buettner, Gabriela Alexe, Björn Häupl, Federico Comoglio, Constanze Schneider, Carmen Doebele, Dominik C. Fuhrmann, Sebastian Wagner, Elisa Donato, Carolin Andresen, Anne C. Wilke, Alena Zindel, Dominique Jahn, Bianca Splettstoesser, Uwe Plessmann, Silvia Münch, Khali Abou-El-Ardat, Philipp Makowka, Fabian Acker, Julius C. Enssle, Anjali Cremer, Frank Schnütgen, Nina Kurrle, Björn Chapuy, Jens Löber, Sylvia Hartmann, Peter J. Wild, Ilka Wittig, Daniel Hübschmann, Lars Kaderali, Jürgen Cox, Bernhard Brüne, Christoph Röllig, Christian Thiede, Björn Steffen, Martin Bornhäuser, Andreas Trumpp, Henning Urlaub, Kimberly Stegmaier, Hubert Serve, Matthias Mann, and Thomas Oellerich



D

Biological processes listed on the y-axis (from top to bottom):

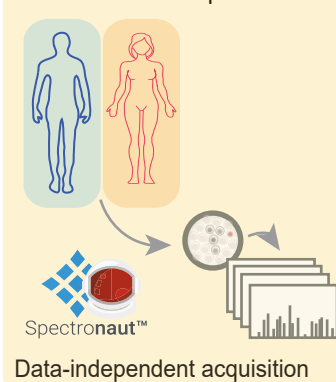
- MYELOID LEUKOCYTE MEDIATED IMMUNITY
- CELL ACTIVATION INVOLVED IN IMMUNE RESPONSE
- MYELOID LEUKOCYTE ACTIVATION
- UBIQUITIN LIKE PROTEIN TRANSFERASE ACTIVITY
- SPECIFIC GRANULE
- AZUROPHIL GRANULE
- PHAGOCYTOSIS
- ACTIN FILAMENT BASED PROCESS
- SECRETORY GRANULE MEMBRANE
- PLATELET DEGRANULATION
- BLOOD MICROPARTICLE
- WOUND HEALING
- RESPONSE TO WOUNDING
- COAGULATION
- MRNA PROCESSING
- RNA SPLICING
- RNA SPLICING VIA TRANSESTERIFICATION
- RIBONUCLEOPROTEIN COMPLEX BIOGENESIS
- MRNA BINDING
- ORGANELLE INNER MEMBRANE
- MITOCHONDRIAL ENVELOPE
- MITOCHONDRIAL MATRIX
- MITOCHONDRIAL PROTEIN COMPLEX
- MITOCHONDRIAL TRANSLATION

Legend:

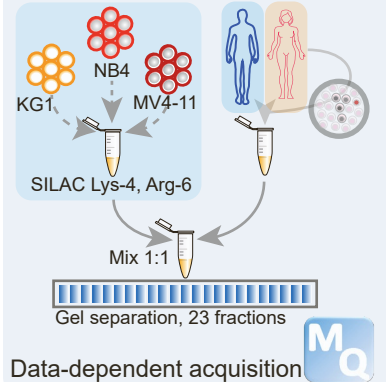
- GeneRatio: 0.05, 0.10, 0.15, 0.20 (dot sizes)
- Adjusted P-Value: 0.01 (red), 0.02 (pink), 0.03 (purple), 0.04 (blue) (dot colors)

Clusters on the x-axis: Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5, Cluster 6

Label-free Mass Spec



SUPER-SILAC-based Mass Spec



MitoTracker Green MFI

ns

Non-Mito C-Mito

Group	Min (%)	Q1 (%)	Median (%)	Q3 (%)	Max (%)
Non-Mito	15	22	45	52	95
C-Mito	25	30	55	68	85

G

C-Mito Non-Mito




Figure S1: Overview of the study design, discovery cohort, the mutational landscape and proteomic subclusters, related to Figure 1

A Available molecular data for the 177 AML patients of the discovery cohort. Rows correspond to molecular data layers, columns to patients. Missing molecular data for a patient in a data layer is shown in red. **B** Patients either belonged to a discovery cohort (n=177) or a validation cohort (n=75). **C** OncoPrint plot showing the most recurrent somatic mutations in 177 AML patients of the discovery cohort. Rows correspond to genomic variants, columns correspond to patients. Variants are ranked by mutations frequency (left), patients are split according to proteomic cluster assignment (bottom). Alterations is color-coded (right). Frequency and type of alterations for each individual sample are shown at the top. **D** Enrichment plot characterizing the six proteomic clusters using a pathway analysis of differentially upregulated proteins between each cluster and the remaining patients. This reveals mitochondrial terms as distinct features of C1 and C2. For each cluster significantly enriched GO and REACTOME pathways are shown ($P < 0.05$). Adjusted p values (Benjamini & Hochberg correction) are color-coded and relative gene set sizes are encoded in symbol size. **E** Mitochondrial content of patient samples (n C-Mito = 8, n non-Mito = 9) measured by flow cytometry using MitoTracker Green (Wilcoxon rank sum-test). MitoTracker Green mean fluorescent intensity was quantified in the $SSC^{low}/CD45^{dim}$ blast gate. **F** Ki67 staining of bone marrow blasts (%) comparing C-Mito (n = 15) and non-Mito (n = 53) patients (left, Wilcoxon rank sum-test). **G** Representative Ki67 staining of C-Mito (left) and non-Mito (right) bone marrow biopsies (scale bar indicates 100 μm). All boxplots in the figure are defined as follows: middle line corresponds to the median; the lower and upper hinges correspond to first and third quartiles, respectively; the upper whisker extends from the hinge to the largest value no further than 1.5 \times the interquartile range (or the distance between the first and third quartiles) from the hinge and the lower whisker extends from the hinge to the smallest value at most 1.5 \times the interquartile range of the hinge. Data beyond the end of the whiskers are called 'outlying' points and are plotted individually if not stated otherwise. ns $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

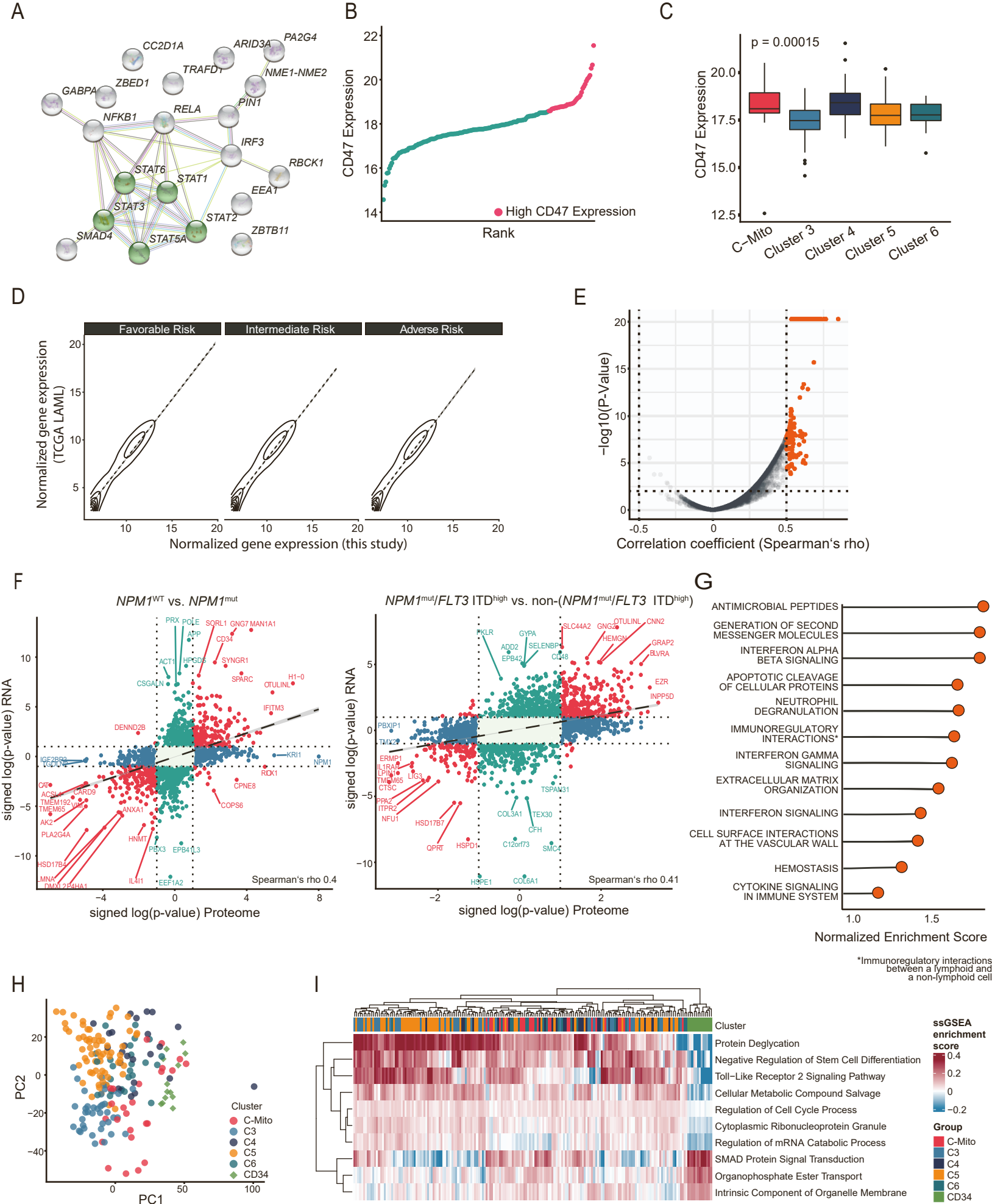


Figure S2: CD34 protome, the transcriptomic landscape, transcriptome-proteome correlations and extended use of the dataset, related to Figure 1

A StringDB-network of the transcription factor cluster in Figure 1D (marked 1) reveals members of STAT family of transcription factors to have a joined regulation within this cluster. **B** Rank plot for CD47 expression, patients with high expression (> third quartile) are color-coded in red. **C** Expression of CD47 per proteomic cluster (Kruskal-Wallis-test). **D** Gene-wise correlation analysis (Spearman rank correlation) of the transcriptome of the discovery cohort (177 patients) and the TCGA AML dataset stratified by cytogenetic risk group. X- and y-axis show normalized expression in the discovery cohort and TCGA AML. For all risk groups correlation coefficients (Spearman's rho) of 0.93 indicate a high comparability. **E** Volcano plot depicting the proteome-transcriptome correlation of $-\log_{10}$ adjusted p value and correlation coefficient (Spearman's rho; Benjamini & Hochberg correction). Proteome-transcriptome correlation ranges from weakly negative to moderately positive with a general modest to moderate positive correlation (left). **F** Scatter plot comparing differentially expressed proteins (x-axis) and genes (y-axis) in *NPM1* mutated vs. wild-type (**left**) and *NPM1*^{mut}/*FLT3*-ITD^{high} (VAF ≥ 0.5) co-mutated vs. non-(*NPM1*^{mut}/*FLT3*-ITD^{high}, **right**) patients. Blue = signed log(p-value) proteome ≥ 1 and signed log(p-value) transcriptome < 1 , red = signed log(p-value) proteome and transcriptome ≥ 1 , green = signed log(p-value) transcriptome ≥ 1 and signed log(p-value) proteome < 1 . **G** Gene Set Enrichment Analysis (GSEA) on the proteome-transcriptome Spearman correlation coefficient. For all significantly enriched terms the normalized enrichment score (NES; normalized to mean enrichment of random samples of the same size) is shown on the X-axis (right). **H** PCA plot after removing batch-induced variability based on the surrogate variable illustrating the inter-group and inter-samples differences in an integrated manner. **I** Single-sample GSEA (ssGSEA) of the batch-corrected proteome including the AML discovery and CD34 cohort identifies differentially regulated pathways.

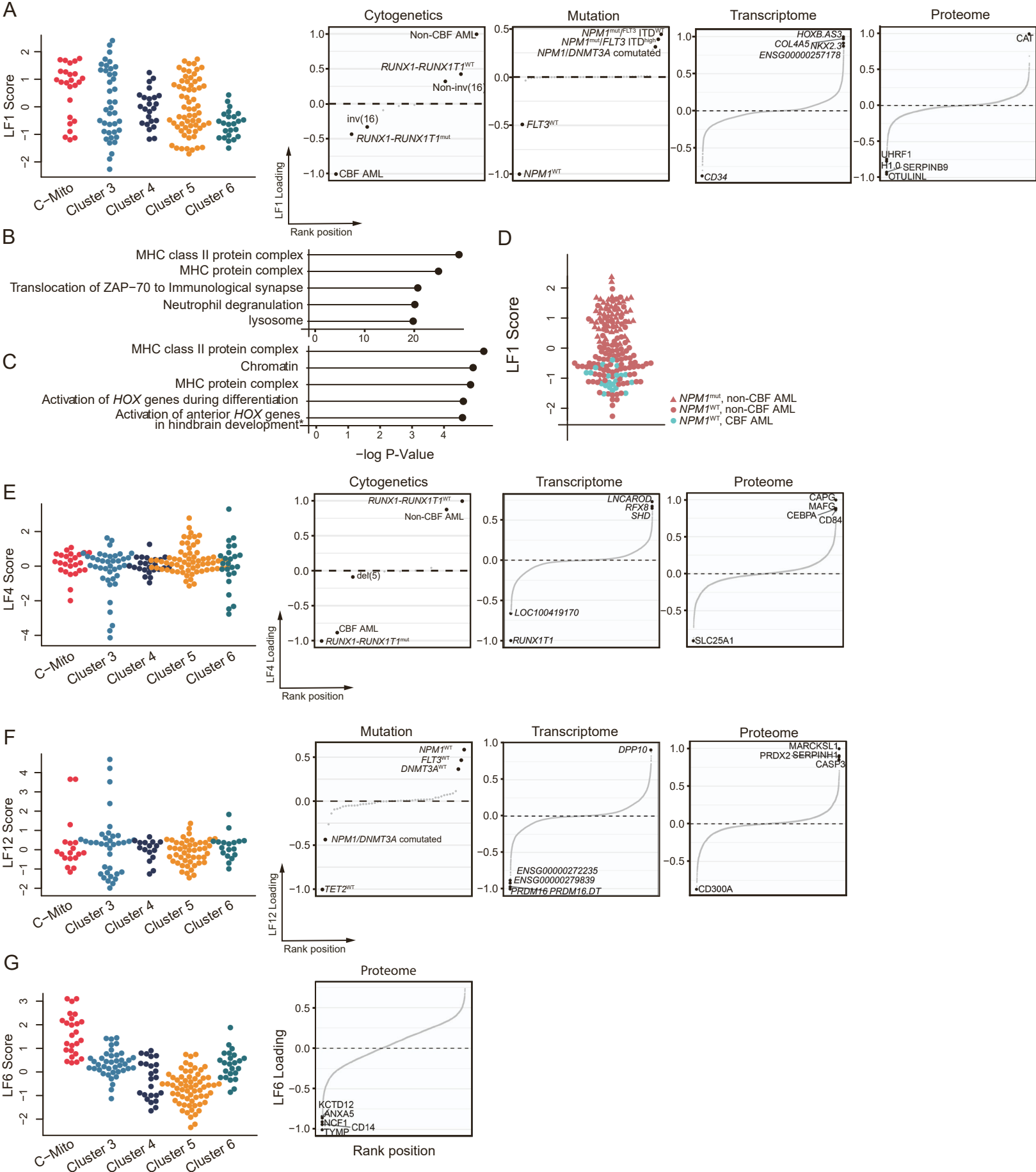


Figure S3: Extended characterization of MOFA Latent Factors (LF), related to Figure 2

Extended characterization of LF1 (A-D), LF4 (E), LF12 (F) and F6 (G) A LF1 score (y-axis) stratified by proteomic cluster (x-axis), where each symbol corresponds to a patient, and data-view specific factor weights for cytoGenetics, mutations, transcriptome and proteome (from left to right). B Pathway analysis of the proteome-specific weights for LF1 (top 5 terms with -log10 adjusted p-value, Benjamini & Hochberg correction). C Pathway analysis of the transcriptome-specific weights for LF1 (top 5 terms with -log10 adjusted p-value, Benjamini & Hochberg correction). D Beeswarm plot of LF1 scores, where each symbol corresponds to a patient. Color encodes CBF status and shape encodes *NPM1* mutation status, illustrating that LF1 separates patients with *NPM1* mutation status from patients with *NPM1* wildtype status. Since all CBF AML patients have *NPM1* wildtype status, LF1 also separates CBF AML patients from patients with *NPM1* mutation status. E LF4 score (y-axis) stratified by proteomic cluster (x-axis), where each symbol corresponds to a patient, and data-view specific factor weights for cytoGenetics, transcriptome and proteome (from left to right). F Score for LF12 (y-axis), stratified by proteomic cluster (x-axis), where each symbol corresponds to a patient, and data-view specific factor weights for mutations, transcriptome and proteome (from left to right). G LF6 score (y-axis) stratified by proteomic cluster (x-axis), where each symbol corresponds to a patient, and data-view specific factor weights for the proteome.

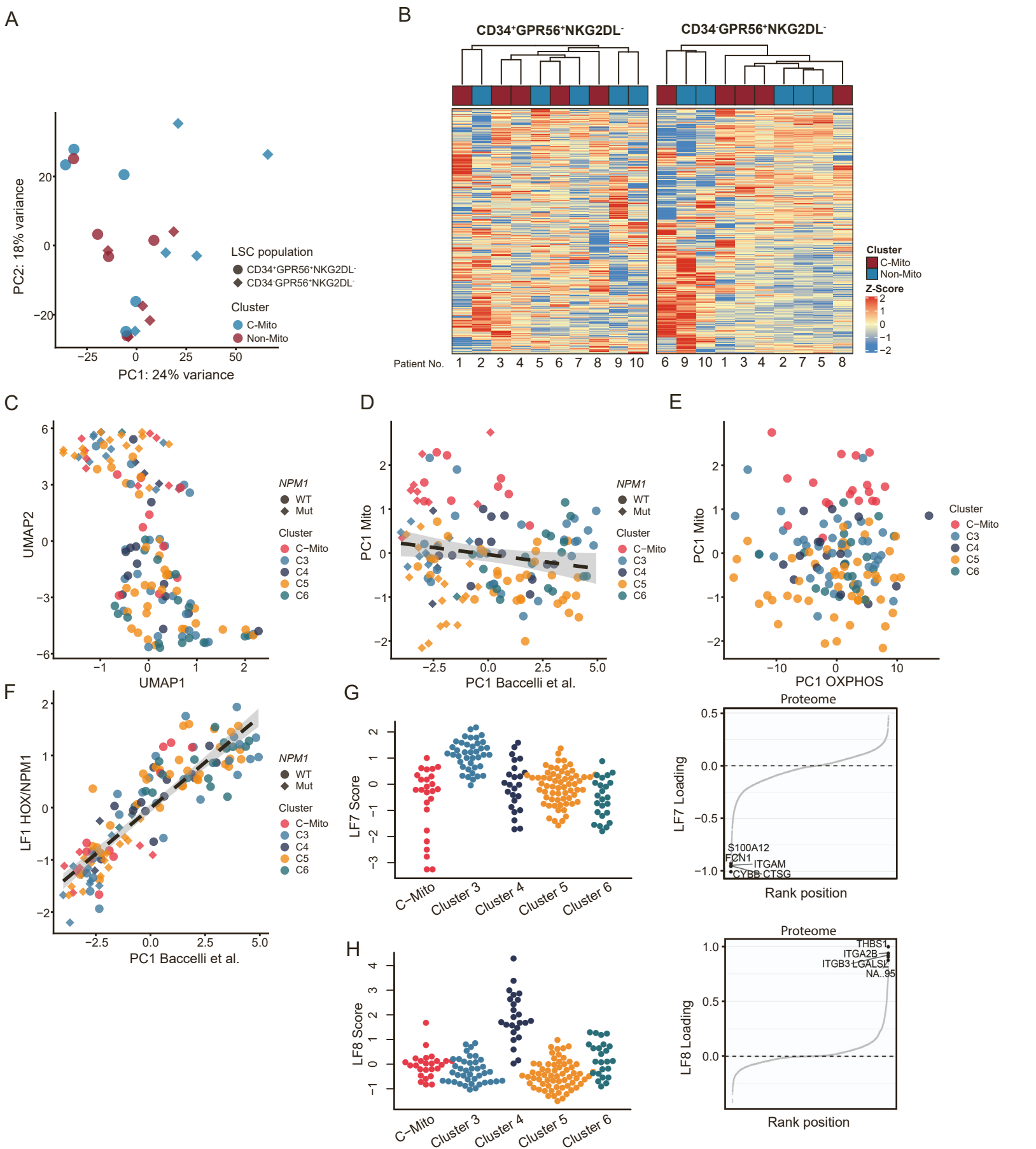


Figure S4: LSC transcriptome, Baccelli et al. genetic signature, related to Figure 2

A Scatter plot of the 1st and 2nd PC derived from the gene expression profile of stem cell subpopulations (circle = CD34⁺GPR56⁺NKG2DL⁻; diamond = CD34⁺GPR56⁺NKG2DL⁺) of 10 discovery cohort patients (n C-Mito = 5, n non-Mito = 5). **B** Heatmap depicting the patient-wise expression of genes related to mitochondria as defined by the human Mitocarta for the CD34⁺ (left) and CD34⁺ (right) stem cell population. Each column is a patient, each row is a gene. Annotated at the top is the proteomic C-Mito classification. Patients are clustered hierarchically based on the Euclidean distance. **C** UMAP of the discovery cohort based on the transcriptomic signature by Baccelli et al. **D** Scatter plot showing a weak correlation between PC1 of the mitochondrial proteome and PC1 scores calculated based on the transcriptomic signature by Baccelli et al. with a correlation coefficient (Spearman's rho) of 0.17. **E** Scatter plot depicting PC1 of the OXPHOS transcriptome vs. the PC1 of the mitochondrial proteome. **F** Scatter plot depicting the PC1 scores calculated based on the transcriptomic signature by Baccelli et al. and LF1; Spearman's rho 0.89). *NPM1* mutation status is coded by shape, proteomic cluster assignments are color-coded. **G,H** LF7 and LF8 score respectively (y-axis) stratified by proteomic cluster (x-axis), where each symbol corresponds to a patient, and data-view specific factor weights for the proteome.

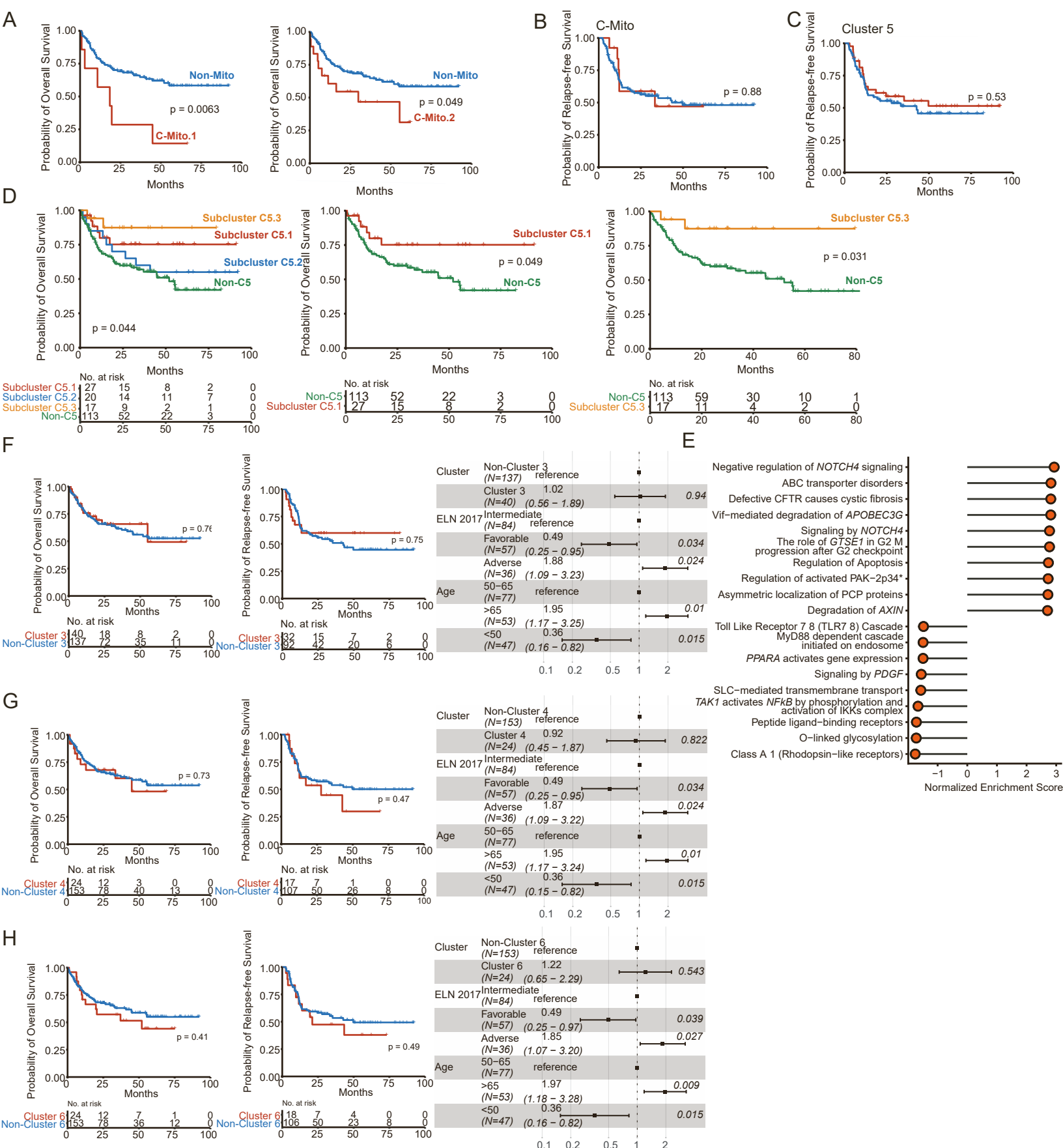


Figure S5: Extended clinical endpoint analyses for the proteomic clusters, related to Figure 3

A Kaplan-Meier model for overall-survival of C-Mito subcluster 1 (C-Mito.1, left) and 2 (C-Mito.2, right) compared to non-Mito patients. **B** Kaplan-Meier model of relapse-free survival for C-Mito compared to non-Mito patients of the discovery cohort. **C** Kaplan-Meier model of relapse-free survival for C5 compared to non-C5 patients of the discovery cohort. **D** Kaplan-Meier model for overall-survival of C5 subcluster 1-3 (left), C5.1 (center) and C5.3 (right) compared to non-C5 patients. **E** Significantly enriched (FDR < 0.01) gene-sets for C5.3 vs. non-C5 (*Regulation of activated PAK-2p34 by proteasome mediated degradation). **F-H** Kaplan-Meier models for overall- and relapse-free survival for C3, C4, C6 and non-C3, non-C4 and non-C6 patients respectively of the discovery cohort (left). Forest plot based on hazard ratios (HRs) of a multivariate Cox regression analysis for the overall survival of patients in the discovery cohort. Covariates used in the Cox regression are shown in the leftmost column. The second column lists all modelled levels of the covariates, with the top-most level being the reference level for each covariate. HR estimates with 95% confidence intervals relative to these references are shown in the third column and visualized as boxes and horizontal lines respectively in the fourth column. The dotted vertical line marks a HR of 1. p-values indicating significance are shown on the right (right).

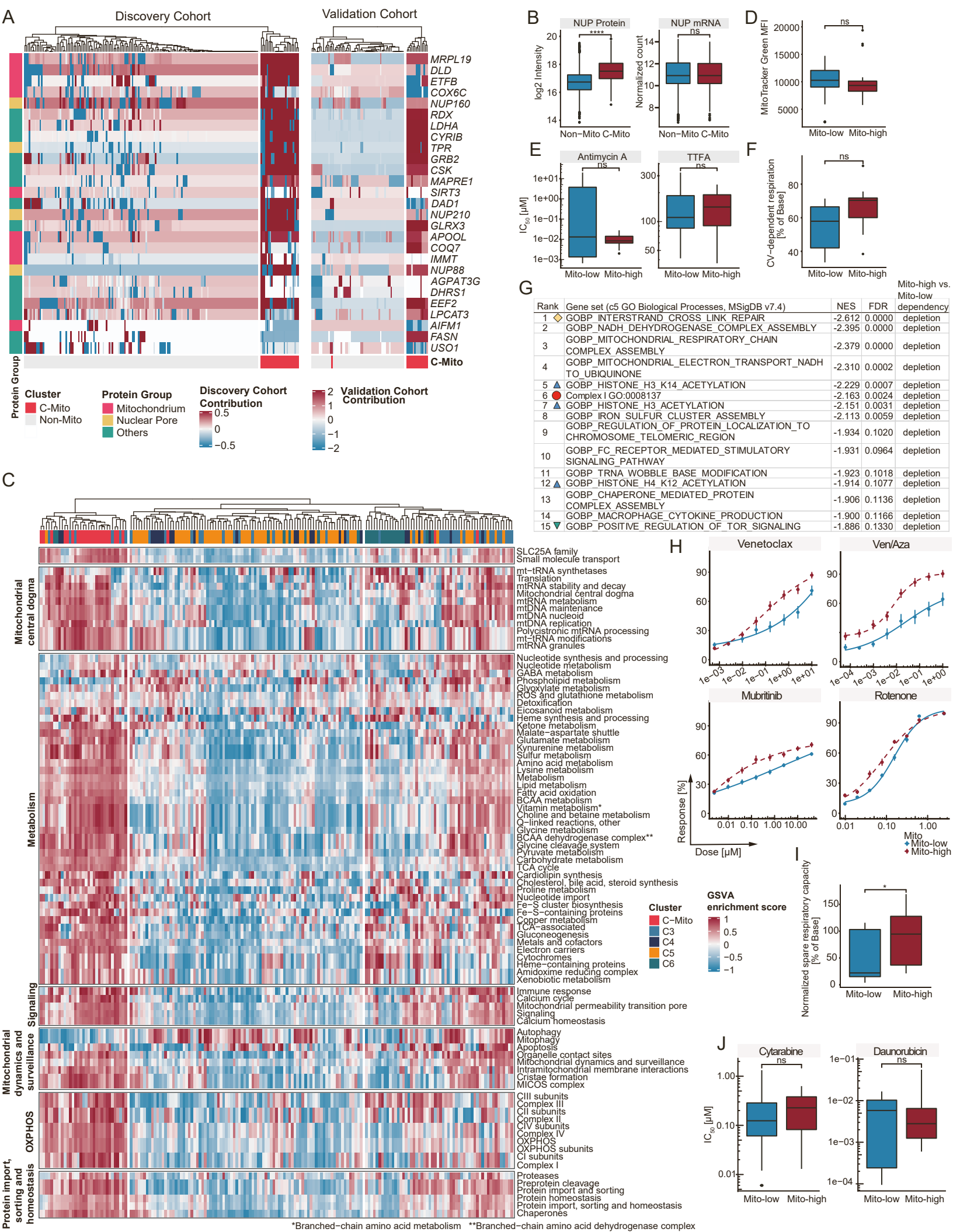


Figure S6: Contribution plot of the signature proteins used to classify C-Mito in the validation cohort, Mitocarta 3.0 heatmap, related to Figures 4, 5 and 6

A Rows are the signature proteins used for classification of C-Mito, columns are samples (177 patients discovery cohort on the left, 75 patients validation cohort on the right) and colors code the contribution per protein used to classify the individual patient into C-Mito or non-Mito (white denotes missing values in the unimputed dataset; we omitted NUP155 since this protein had zero contribution to predictions in the validation cohort). Hierarchical clustering of the contributions recapitulates the classification generated by the non-linear gradient boosting-based classifier. Clustering was performed with complete linkage of Euclidean distances. The annotation on the left shows protein group memberships as defined by UniProt keywords. For visualization, protein-wise Z-scaling was done independently for both datasets (discovery and validation cohort). **B** Expression of nuclear pore subunits in the proteome is significantly higher in Mito patients (log2 intensity, $p < 0.001$) while there is no significant difference on the gene level (normalized expression). **C** Heatmap depicting the gene-set variation analysis (GSVA) enrichment scores based on mitochondrial pathways defined by the human Mitocarta 3.0. Columns are discovery cohort patients, rows are mitochondrial gene-sets. Color-coded at the top are the proteomic cluster assignments. Hierarchical clustering (complete linkage) based on Euclidean distance is used to cluster the columns. The MitoPathways top level hierarchy terms are used to group the rows. **D** Mitochondrial content of AML cell lines (n Mito-high = 8, n Mito-low = 9, measured in triplicates) measured by flow cytometry using MitoTracker Green (Wilcoxon rank sum-test). **E** The half-maximum inhibitory concentration (IC_{50}) of Mito-high (n=8) and Mito-low AML cell lines (n=8) is shown. Mito-high and Mito-low AML cell lines have similar sensitivity towards OXPHOS complex II (TTFA) and III (Antimycin A) inhibitors (Wilcoxon rank sum test). **F** Complex V-dependent respiration is assessed by comparing the oxygen consumption rate (OCR) of Mito-high (n=8) and Mito-low (n=6) cell lines before and after oligomycin treatment. No significant difference is observed (Wilcoxon rank sum test). **G** Top 15 significantly enriched gene-sets (GO biological processes) for the genome-wide differential CERES dependency score in Mito-high vs Mito-low cell lines available in the CRISPR (Avena) 21Q2 data. Significance: $abs(normalized\ enrichment\ score(NES)) > 1.3$, $p\text{-value} < 0.1$. OXPHOS. Symbols are used to identify the terms in the volcano plot Figure 6B. **H** Dose-response models comparing the sensitivity of Mito-high (n = 9) vs. Mito-low (n = 10) cell lines to venetoclax, venetoclax + azacitidine, mubritinib and rotenone. **I** Box plot depicting the difference in normalized spare respiratory capacity (% of base respiration) between Mito-high (n = 8) and Mito-low (n= 6) AML cell lines ($p = 0.043$, Wilcoxon rank sum test). **J** Box plot depicting the IC_{50} for cytarabine (left) and daunorubicin (right) in Mito-high (n=9) and Mito-low (n=10) cell lines (Wilcoxon rank sum test).

All boxplots in the figure are defined as follows: middle line corresponds to the median; the lower and upper hinges correspond to first and third quartiles, respectively; the upper whisker extends from the hinge to the largest value no further than $1.5 \times$ the interquartile range (or the distance between the first and third quartiles) from the hinge and the lower whisker extends from the hinge to the smallest value at most $1.5 \times$ the interquartile range of the hinge. Data beyond the end of the whiskers are called 'outlying' points and are plotted individually if not stated otherwise. ns $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.