SUPPLEMENT

Characterization of myeloproliferative neoplasms based on genetics only and prognostication of transformation to blast phase

Wencke Walter, Niroshan Nadarajah, Stephan Hutter, Heiko Müller, Claudia Haferlach, Wolfgang Kern, Torsten Haferlach, Manja Meggendorfer

MLL Munich Leukemia Laboratory, Munich, Germany

Table of content

1. Patients cohorts and methods

1.1. Patients' cohorts

Training and validation cohorts for machine learning algorithm for disease classification: the model was built based on a cohort of 355 well characterized samples diagnosed with either CML (n=106), PMF (n=97), PV (n=75), or ET (n=77) based on morphology and *JAK2*, *MPL*, *CALR* mutation status following WHO criteria (cohort 1). Histopathology was not performed in house. The diagnostic relevant criteria were based on information provided by the sending physicians or provided histopathological reports. The median age was 65 years (range: 16-87 yrs) and male/female ratio was 1.5 (211/144). Whole genome and transcriptome sequencing was performed for all samples and the mutation status of 73 genes, recurrently mutated in myeloid malignancies, assessed. Further, in 323/355 cases also cytogenetics was available. In addition, in 314/355 patients' blood counts in terms of leukocytosis (>15,000/µL), erythrocytosis (Hb >16.5 g/dL in men and >16 g/dL in women), and thrombocytosis (platelet counts ≥450 $x10^9$/L) were available. For validation of the final model, we used a completely independent cohort, based on recent cases analyzed in routine diagnostics with an NGS myeloid gene panel with 28 genes listed in table 1 (1) covering all genes used by the final model (cohort 2). The validation cohort comprised in total 109 patients: 37 CML, 24 ET, 29 PV and 19 PMF. The median age was 66 years (range: 23-85 yrs) and male/female ratio was 1.2 (60/49).

MPN-NOS cohort: Whole genome and transcriptome sequencing was performed for 55 cases diagnosed with MPN-NOS (cohort 3). The median age was 70 years, range: 37-91 yrs; male/female ratio: 1.2 (30/25). The mutation status of 73 genes, recurrently mutated in myeloid malignancies, was assessed and used for genetic classification by the developed model.

Cohorts for genetic characterization of MPN progressing to blast phase: the total cohort consisted of 19 patients (11 male, 8 female, ratio: 1.4) investigated at both time points of MPN-CP and progression to blast phase (MPN-BP) by whole exome sequencing (cohort 4). The cohort included PMF (n=9), ET (n=2), and PV (n=3). In 5 patients a clear sub-classification was not possible. The median age at MPN-CP was 72 years (range: 61-89 yrs), median time

between chronic phase and progression to blast phase was 33 months (range: 8-82 months). Cytogenetics was available in 15/19 patients at both time points. The mutation pattern of MPN patients progressing to blast phase was compared to a random MPN cohort of 34 patients representing all three MPN subtypes (PMF, n=13; ET, n=11; PV, n=10) and showing no progression to blast phase within an median observation period of 86 months (range: 10-233 months, cohort 5). The male/female ratio was 1.3 (19/15) and the median age was 64 years (range: 28-83 yrs). These patients were investigated by a targeted NGS myeloid panel covering 28 genes (table 1), described previously (1).

1.2. Genes investigated by panel sequencing

Table 1: 28 analyzed genes by panel sequencing

ASXL1	*EZH2*	*KIT*	*SETBP1*
BCOR	*FLT3*-TKD	*KRAS*	*SF3B1*
BRAF	*GATA1*	*MPL*	*SRSF2*
CALR	*GATA2*	*NPM1*	*TET2*
CBL	*IDH1*	*NRAS*	*TP53*
DNMT3A	*IDH2*	*PHF6*	*U2AF1*
ETV6	*JAK2*	*RUNX1*	*WT1*

1.3. Whole exome sequencing (WES)

DNA was extracted from bone marrow or peripheral blood after lysis of erythrocytes using the MagNA Pure 96 Instrument (Roche Diagnostics, Mannheim, Germany). Whole exome sequencing was performed by the TruSeq DNA Exome Library Prep (Illumina, San Diego, CA, USA), starting with 150 ng of genomic DNA. The coding regions (exome) were enriched by the xGen Exome Research Panel v1 using a hybridization capture workflow (IDT Integrated DNA Technologies, Coralville, IA, USA), targeting 19,433 genes. All libraries were sequenced with 100 bp paired-end reads on a NovaSeq6000 (Illumina) with a median coverage of 257x. The NovaSeq Fastq files were aligned to the GRCh37 reference genome using the iSAAC Aligner (03.16.02.20) and the variant calling for small nucleotide variants (SNVs) was performed with Pisces (5.1.3.60) with a 1% variant allele frequency (VAF) cutoff and annotated with Nirvana (1.3.5.2) and VEP. For a set of 165 genes (Table 1), known to be mutated in myeloid and lymphoid neoplasms, all variants were considered that were flagged as PASS and within (+/- 2 CCDS (consensus coding sequence)) exons. For the rest of the exome, the following filtering

strategy was applied for SNVs: VAF >5%, gnomAD v2.2.1 global frequency cutoff at 0.005% and either an in-house consensus functional prediction score >0.5 (2) or a protein truncating mutation. Copy number variation (CNV) and copy neutral loss of heterozygosity (CN-LOH) calling from exome data were performed by cnvkit (3), with coding sequences defined as on-target regions and the remainder as off-target regions. The plots for each patient were also manually reviewed to correct for events potentially missed by the algorithm.

Table 2: 165 analyzed genes by whole exome sequencing

ABL1	*BRCC3*	*CEBPA*	*DIS3*	*FAT4*	*HBB*	*KDM6A*	*MYC*	*PRPF8*	*SMC1A*	*TP53*
ABL2	*BRD1*	*CHEK2*	*DNMT3A*	*FBXW7*	*HIF1A*	*KDR*	*MYD88*	*PTEN*	*SMC3*	*TRAF3*
APC	*BRD2*	*CRBN*	*EGLN1*	*FGR*	*HIF1AN*	*KIT*	*NF1*	*PTPN11*	*SRC*	*TYK2*
ARID1A	*BRD3*	*CREBBP*	*EGLN2*	*FLT3*	*HIF3A*	*KLF2*	*NFKBIE*	*PTPRD*	*SRSF2*	*U2AF1*
ASXL1	*BRD4*	*CSF1R*	*EGLN3*	*FLT3ITD*	*ID3*	*KLHL6*	*NOTCH1*	*RAD21*	*STAG2*	*U2AF2*
ASXL2	*BRDT*	*CSF3R*	*EP300*	*FOXO1*	*IDH1*	*KMT2A*	*NOTCH2*	*RB1*	*STAT3*	*UBR5*
ATM	*BTK*	*CSNK1A1*	*EPAS1*	*FYN*	*IDH2*	*KMT2D*	*NPM1*	*RET*	*STAT5A*	*VHL*
ATRX	*CALR*	*CTCF*	*EPO*	*GATA1*	*IKBKB*	*KRAS*	*NRAS*	*ROS1*	*STAT5B*	*VPS45*
BCL2	*CARD11*	*CUX1*	*EPOR*	*GATA2*	*IKZF1*	*LCK*	*NTRK1*	*RPS15*	*SUZ12*	*WHSC1*
BCOR	*CBL*	*CXCR4*	*ETNK1*	*GFI1B*	*IL2RG*	*LRP1B*	*OS9*	*RUNX1*	*TBL1XR1*	*WT1*
BCORL1	*CCND1*	*DAPK3*	*ETV6*	*GNAS*	*JAK1*	*MAP2K1*	*PHF6*	*SETBP1*	*TCF3*	*XPO1*
BHLHE41	*CD79A*	*DDX3X*	*EZH2*	*GNB1*	*JAK2*	*MAPK1*	*PIGA*	*SF1*	*TET2*	*ZBTB7A*
BIRC3	*CD79B*	*DDX41*	*FAM46C*	*GPR98*	*JAK3*	*MEF2B*	*PLCG2*	*SF3A1*	*TLR2*	*ZMYM3*
BPGM	*CDH23*	*DDX54*	*FANCL*	*HBA1*	*KDM1A*	*MPL*	*POT1*	*SF3B1*	*TNFAIP3*	*ZNF197*
BRAF	*CDKN2A*	*DHX29*	*FAS*	*HBA2*	*KDM5A*	*MYBBP1A*	*PPM1D*	*SH2B3*	*TNFRSF14*	*ZRSR2*

1.4. Whole genome sequencing (WGS)

WGS libraries were prepared from 1µg of DNA with the TruSeq PCR free library prep kit following the manufacturer's recommendations (Illumina, San Diego, CA, USA) and 2x150bp paired-end sequences were generated on a NovaSeq 6000 or HiSeqX instrument with 100x coverage (Illumina, San Diego, CA, USA). Reads were aligned to the human reference genome (GRCh37, Ensembl annotation) using the Isaac aligner (v3.16.02.19) (4) through BaseSpace's WGS app (v5, Illumina, San Diego, CA, USA) with default parameters. Resulting BAM files were used in BaseSpace's Tumor/Normal app (v3) to call single nucleotide variants (SNV) and small indels (<50bp) with Strelka (v2.4.7) (5) and large scale structural variants (SV) with Manta (v0.28.0) (6). As no sample specific normal tissue was available, a so-called unmatched normal was used in its place to reduce technical artefacts and germline calls. For this reason, WGS was performed on gender-matched genomic DNA from a mixture of multiple anonymous donors (Promega, Fitchburg, WI, USA). To further remove potential germline variants, each SNV/small indel was queried against the gnomAD database (v2.1.1) (7) and

variants with global population frequencies >0.05% where excluded. Further analysis was performed on protein-altering and splice-site variants only. CNVs were called with GATK (v4.0.8.1) (8) using the Broad Institute's recommended best practices pipeline. Here, two panels of normals (PON) were used for denoising consisting of 124 female and 191 male samples which presented a normal karyotype during routine diagnostics. CN-LOH was assessed using HadoopCNV (9). *JAK2* Haplotype Calling 46/1 was performed applying the GATK germline calling pipeline to identify 4 associated SNPs (rs3780367, rs10974944, rs12343867, and rs1159782), also called the *JAK2* GGCC-haplotype. Hierarchical clustering of patients' genetic feature vectors using a binary distance metric was performed to identify coincidence patterns of the detected genetic abnormalities.

1.5. Whole transcriptome sequencing

For transcriptome analysis the TruSeq Total Stranded RNA kit was used, starting with 250ng of total RNA, to generate RNA libraries following the manufacturer's recommendations (Illumina, San Diego, CA, USA). 2x100bp paired-end reads were sequenced on the NovaSeq 6000 (Illumina, San Diego, CA, USA) with a median of 50 million reads per sample. Using BaseSpace's RNA-seq Alignment app (v2.0.1) with default parameters reads were mapped with STAR aligner (v2.5.0a) (10) to the human reference genome hg19 (RefSeq annotation). SNVs/small indels were called using Isaac variant caller (v2.3.13) (4). Fusion calling was performed with Manta (v0.29.0) (6), Arriba (v1.2.0) (11) and STAR-Fusion (v1.9.0) (12). Only fusions not present in a panel of 64 healthy control samples, that were called by at least two algorithms and that matched WGS SV calls were considered. Gene specific read abundance was calculated with Cufflinks (v2.2.1) (13).

For transcriptional profile analysis estimated read counts for each gene were normalized applying Trimmed mean of M-values (TMM) normalization method and the resulting log2 counts per million (CPMs) were used. Genes were kept if they were expressed (> 5 CPM) in at least 66% of all samples. Gene expression differences were assessed using the edgeR package (14) with false discovery rate (FDR) correction for multiple testing. Genes with FDR

less than 0.05 and an absolute log2-based fold-change greater than 1.5 were considered differentially expressed (DE). The global gene expression profiles were assessed by principal component analysis.

2. Supplemental figures and tables

2.1. Analyses of gene expression patterns for differentiation of CML, PV, ET, and PMF (cohort 1)

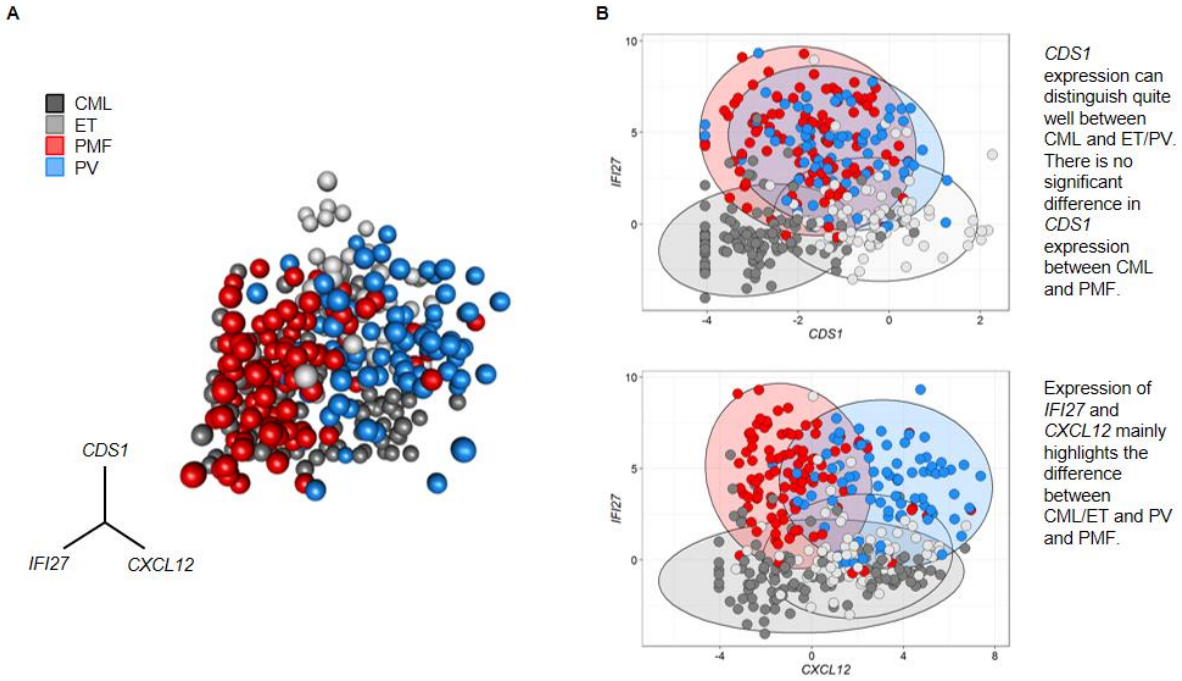

Figure 1: Expression patterns. Illustration of principle component analysis (PCA) showing a clustering by differentially expressed genes *CDS1*, *IFI27*, and *CXCL12* in A) a 3D model and B) separated in two 2D plots.

2.2. Overview on CN-LOH detected in the trainings cohort for genetic characterization of MPN entities (cohort 1)

Table 3: Copy neutral loss of heterozygosity (CN-LOH)

CN-LOH	Entity	Number of cases (n=97)
9p	PV	54
9p	PMF	23
9p	ET	1
9p; 14q	PV	2
9p; 12q	PMF	2
9p; 18q	PMF	1
9p; 22q	PMF	1
1p	PMF	4
1q	PMF	1
4q	PMF	1
7q	PMF	3
10q	CML	1
11q	PMF	1
14q	PMF	1
17p	PMF	1

2.3. Chromosomal aberrations in the trainings cohort used for machine learning approach

Table 4: The 12 most abundant cytogenetic aberrations in the trainings cohort (cohort 1)

Cytogenetic aberration	Number of cases
BCR::*ABL1*	107
gain 1q	4
deletion 5q	5
deletion 7q	4
trisomy 8	11
gain 9p	13
deletion 11q	3
deletion 12p	2
deletion 13q	5
deletion 17p	1
deletion 20q	10
Complex karyotype	5

2.4. Classification of MPN-NOS by the classification model and decision tree

We raised the question if the ML model is able to classify MPN, not otherwise specified (cohort 3, 55 MPN-NOS cases). The model assigned 27% of the patients to PMF, 60% to PV and 13% to ET. By applying the decision tree to the MPN-NOS cases, a similar distribution was observed. The results indicate that MPN-NOS reflect a mixture of genetic profiles of different MPN entities, allowing for a genetically based classification of MPN-NOS patients (figure 2).

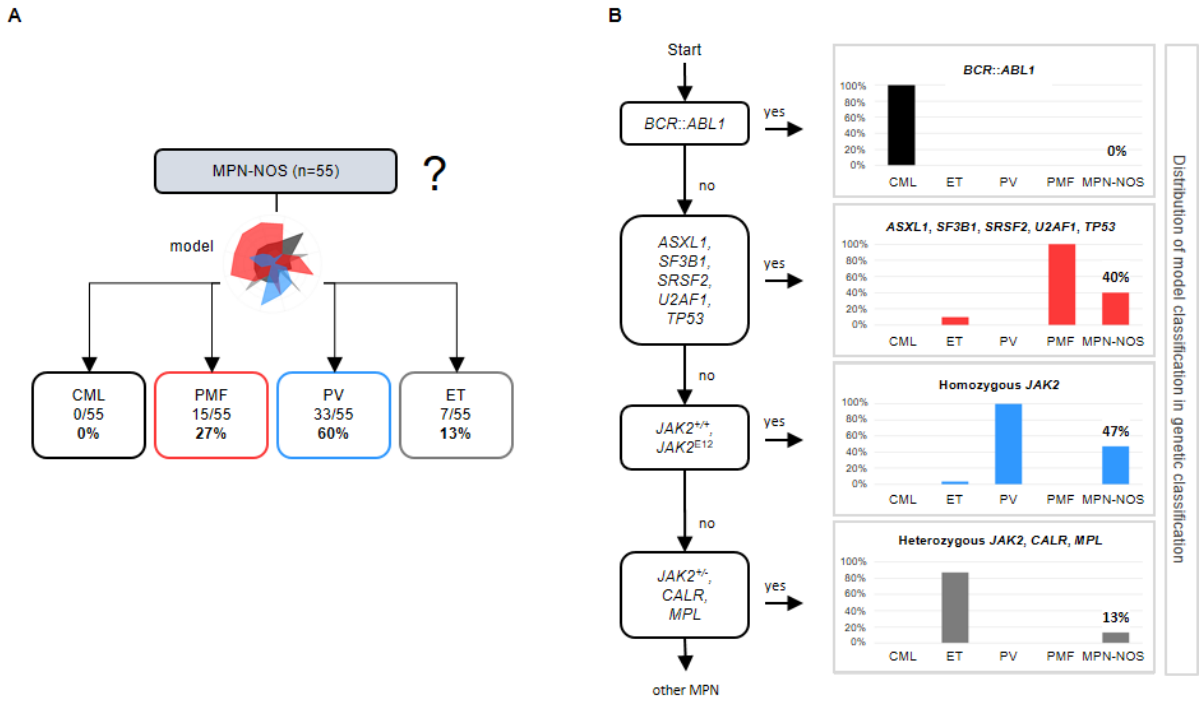

Figure 2: Classification of MPN-NOS (cohort 3). Classification based on A) machine learning model and B) the proposed decision tree. The bars represent the proportion of cases of the validation cohort, as well as MPN-NOS cases, that were classified by the model as CML, PV, ET or PMF and their distribution within the genetic classification.

2.5. Clonal evolution of MPN-CP cases transforming to MPN-BP – mutational and cytogenetic profiles

The tumor phylogeny of the samples was inferred based on the VAF of detected somatic mutations. Mutations with similar VAFs (+/- 5%) were considered to occur in the same clone, including linear and branched evolution models. Clonal evolution of individual patients was visualized using alluvial diagrams, where the indicated genes and aberrations demonstrate clone characteristics and clone width illustrates clone size. The upper part of the diagram shows the clones at MPN-CP, while the lower part shows the clones at MPN-BP. Because bulk-seq data are used in the analysis and the sample purity is largely unknown, the reconstructed clonal substructures are only estimates, and mischaracterized relationships may occur.

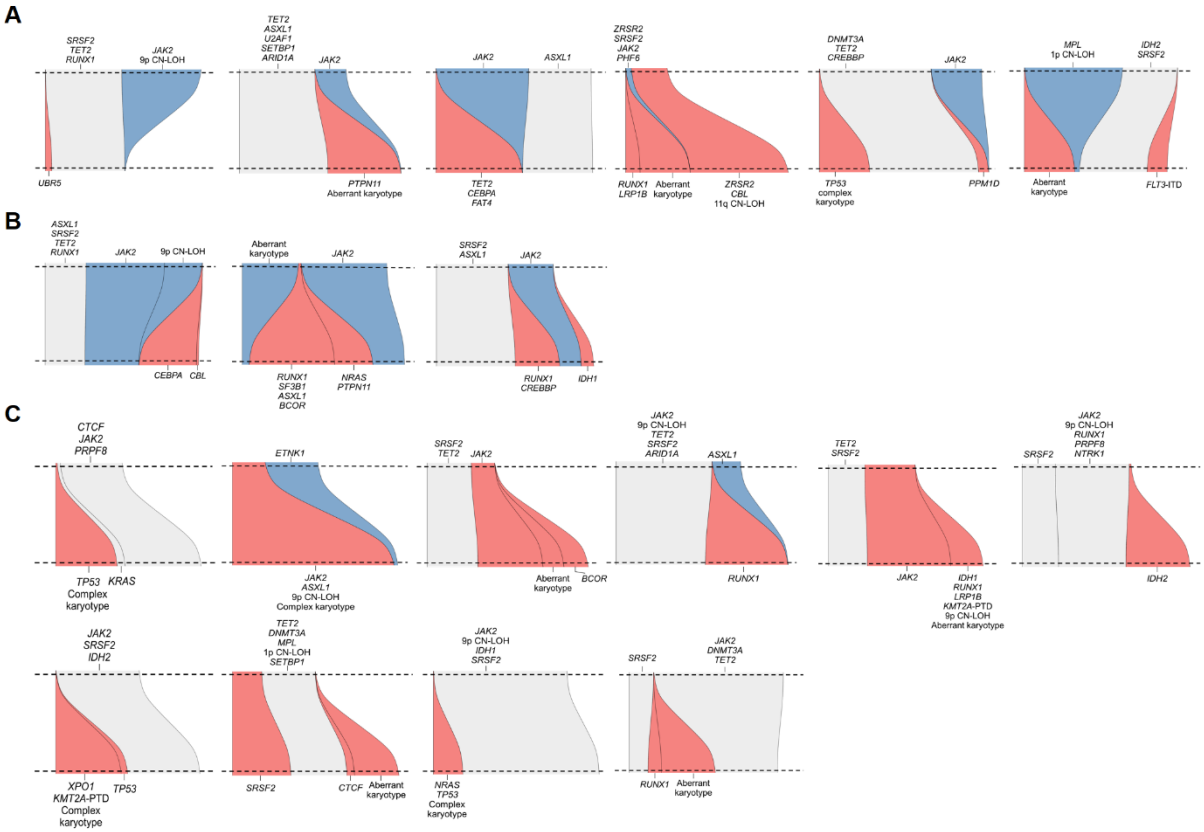

Figure 3: Clonal evolution during progression to MPN-BP (cohort 4). Each chart illustrates the clonal evolution of a single patient. The top of the graph reflects the clones at MPN-CP, while the lower part reflects the clones at MPN-BP. The indicated genes and aberrations show the characteristic of the clone and the width of the clones illustrate the clone size. A red marked clone shows an increasing clone size during progression, while a blue one a decreasing and the grey one a stable clone size. Gene names at the bottom of the graph illustrate gained gene mutations or aberrations. A) Patients losing the MPN-driver mutation during progression. B) Patients showing a decrease of the MPN-driver clone, and C) patients with a stable or increasing MPN-driver gene mutation.

2.6. Clonal evolution of MPN-CP cases transforming to MPN-BP – transcriptomic profiles

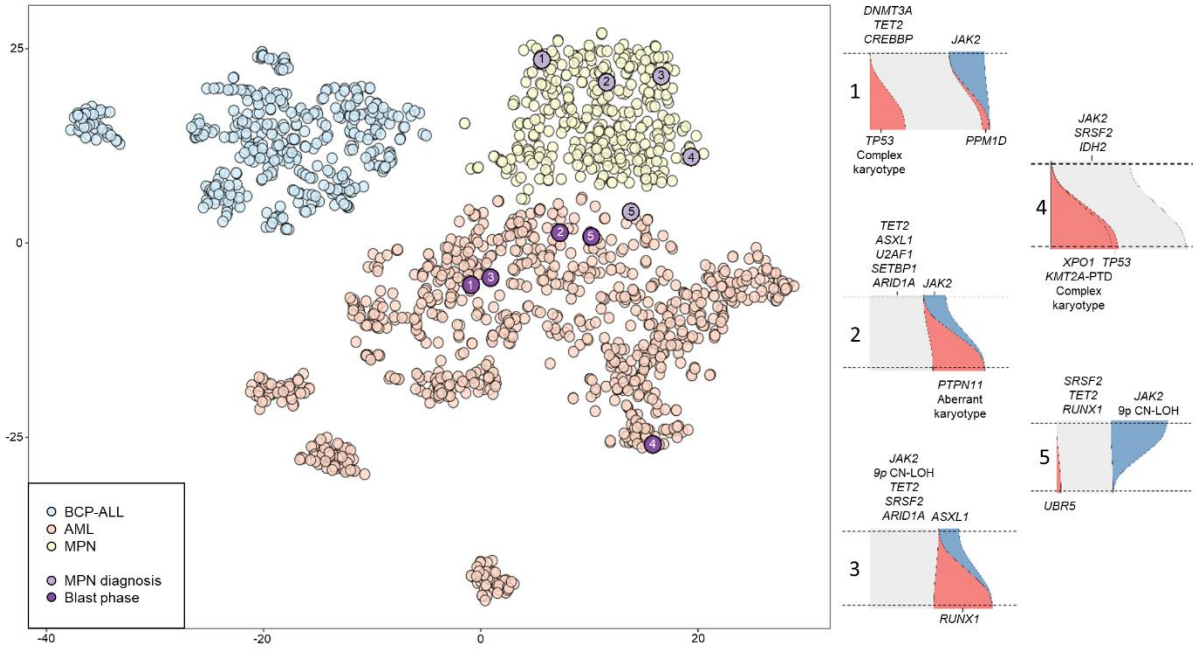

Figure 4: Changing transcriptomic profiles during progression to MPN-BP (cohort 4). The left side shows a UMAP based on the top 15% most variable genes between BCP-ALL (light blue), AML (light orange), and MPN-CP (light yellow) samples. Each dot represents a sample. The paired MPN-CP (light purple) and MPN-BP (dark purple) samples are highlighted with dots of an increased size. On the right side, the clonal evolution patterns of the individual pairs are displayed. The indicated genes and aberrations show the characteristic of the clone and the width of the clones illustrate the clone size. A red marked clone shows an increasing clone size during progression, while a blue one a decreasing and the grey one a stable clone size. Gene names at the bottom of the graph illustrate gained gene mutations or aberrations.

3. References

1. Delic S, Rose D, Kern W, Nadarajah N, Haferlach C, Haferlach T, et al. Application of an NGS-based 28-gene panel in myeloproliferative neoplasms reveals distinct mutation patterns in essential thrombocythaemia, primary myelofibrosis and polycythaemia vera. Br J Haematol. 2016.

2. Hutter S, Baer C, Walter W, Kern W, Haferlach C, Haferlach T. A Novel Machine Learning Based in silico Pathogenicity Predictor for Missense Variants in a Hematological Setting. Blood. 2019;134(Supplement_1):2090.

3. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016;12(4):e1004873.

4. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29(16):2041-3.

5. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591-4.

6. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32(8):1220-2.

7. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

8. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 1 ed: O'Reilly Media, Inc.; 2020.

9. Yang H, Chen G, Lima L, Fang H, Jimenez L, Li M, et al. HadoopCNV: A dynamic programming imputation algorithm to detect copy number variants from sequencing data. bioRxiv. 2017.

10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

11. Uhrig S, Ellermann J, Walther T, Burkhardt P, Fröhlich M, Hutter B, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res. 2021;31(3):448-60.

12. Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol. 2019;20(1):213.

13. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511-5.

14. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.