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 2022 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 (n=87/355). 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. Histopathological reports were available for 6/109 cases. 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). Histopathological reports were available for 11/55 cases. The median age was 70 years, range: 37-91 yrs; male/female ratio: 1.2 (30/25). The mutation status of 73 genes (indicated in table 2), 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. Histopathological reports were available for 4/19 patients. 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. To address which pathways were most frequently affected during MPN-CP and MPN-BP, mutated genes were functionally annotated by three different data bases (3-5). In 15 patients karyotype information was available by chromosome banding analysis, in the remaining 4 patients copy number variation (CNV) and copy neutral loss of heterozygosity (CN-LOH) calling from exome data were performed by cnvkit (6), 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 and 73 (marked in grey) analyzed genes by whole exome and whole genome sequencing, respectively

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) (7) 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) (8) and large scale structural variants (SV) with Manta (v0.28.0) (9). 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) (10) 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) (11) 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 (12). *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) (13) to the human reference genome hg19 (RefSeq annotation). SNVs/small indels were called using Isaac variant caller (v2.3.13) (7). Fusion calling was performed with Manta (v0.29.0) (9), Arriba (v1.2.0) (14) and STAR-Fusion (v1.9.0) (15). 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) (16).

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 (17) 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.

1.6. Machine learning model to stratify MPN patients

The dataset (n = 355, cohort 1) was split into a training and test set (50:50) and a machine learning algorithm was trained to separate the different MPN entities based on molecular markers only. Initially only binary mutation profiles were considered to build 500 models with an accuracy range from 27% - 100% and a median accuracy of 72%. Mutated genes that occurred in more than 60% of the models were kept for the next round of training and expanded by the binary encoded VAF of MPL (<35%/>35%), and JAK2 (<35%/>35%; <60%/>60%). This refinement increased both the median accuracy of all models to 79% and also the number of models with an accuracy >90%. Consideration of the 12 most abundant cytogenetic aberrations (table 4) did not provide further improvement. Therefore, the final model contained only 12 molecular markers: mutation status of *ASXL1*, *BCR*::*ABL1*, *CALR*, *JAK2*, *SF3B1*, *SRSF2*, *TET2*, *TP53*, and *U2AF1* and the binary VAF values *CALR* >35%, *JAK2* >35%, and *JAK2* >60%. The final model stratified patients with an accuracy of 98.3%. If blood counts (in terms of leukocytosis; white blood cell count >15,000/µL; erythrocytosis, Hb >16.5 g/dL in men and >16 g/dL in women; thrombocytosis, platelet counts ≥450 x10^9/L) were considered as well, the accuracy increased slightly (98.9%) with a smaller set of four contributing factors (mutation status of *ASXL1*, *BCR*::*ABL1*, *CALR*, VAF *CALR* >35% beside thrombocytosis, leukocytosis and erythrocytosis). Since the aim was a classification based on genetics only, we stuck to the molecular genetic based model.

2. Supplemental figures and tables

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

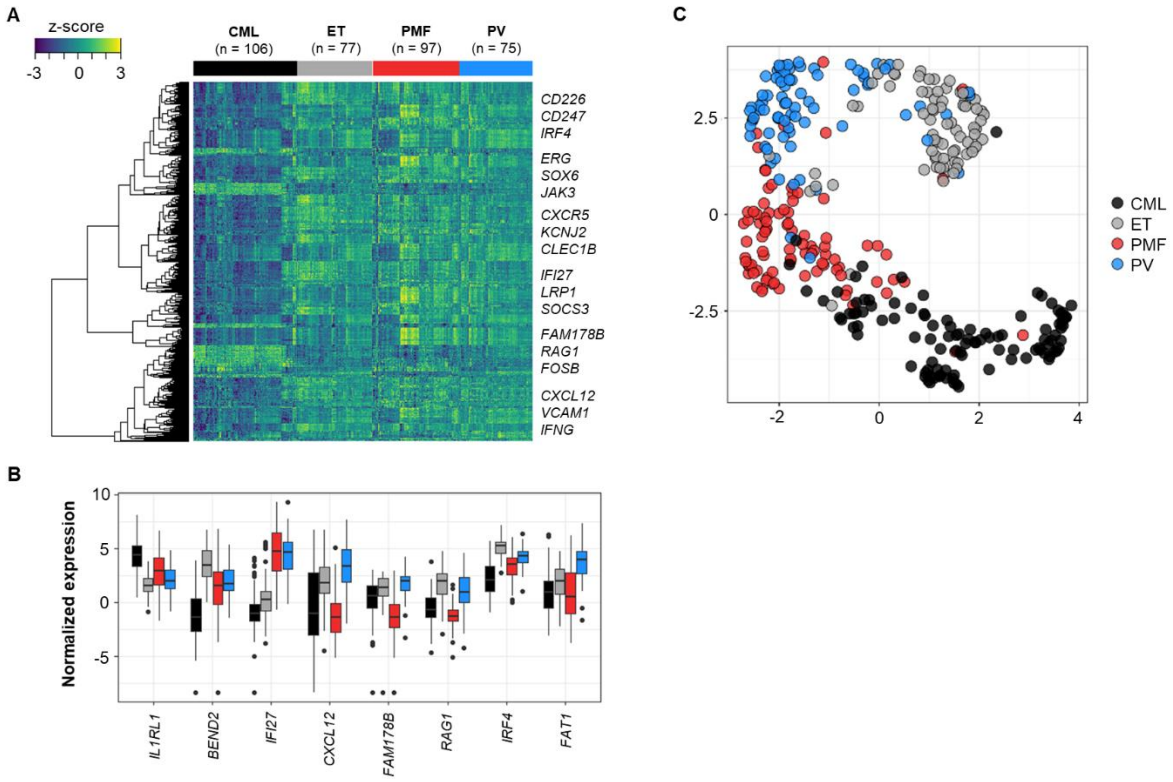

Figure S1: Gene expression profiles of MPN diagnostic groups. A) Heatmap of 670 genes differentially expressed between at least two groups, with high expression levels represented by yellow and low expression levels by dark blue. To the right of the heatmap, selected genes are highlighted. B) Boxplot and C) UMAP of normalized expression of key selected genes to stratify MPN patients.

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. 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.

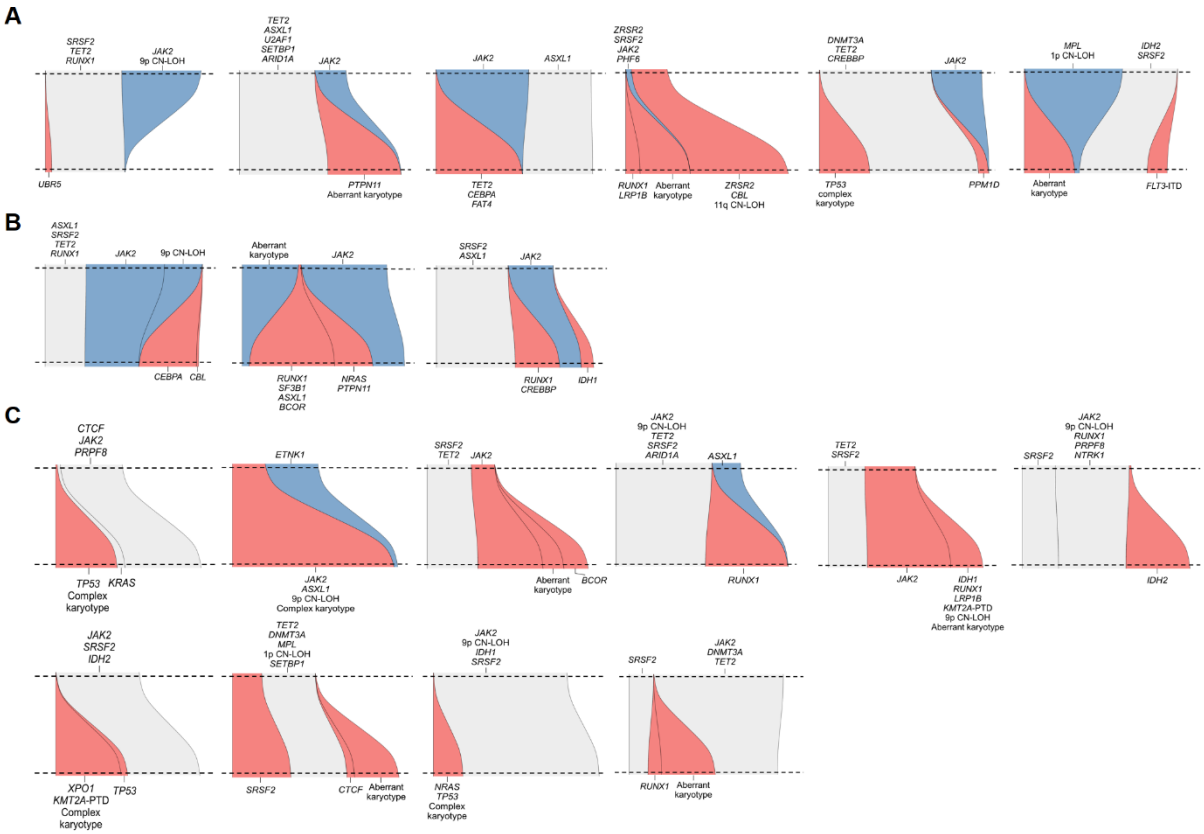

Figure S2: Clonal evolution during progression to MPN-BP (cohort 4). Each chart illustrates the clonal evolution of a single patient. The top of the graph reflects the clones at MPN-CP, while the lower part reflects the clones at MPN-BP. The indicated genes and aberrations show the characteristic of 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.

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

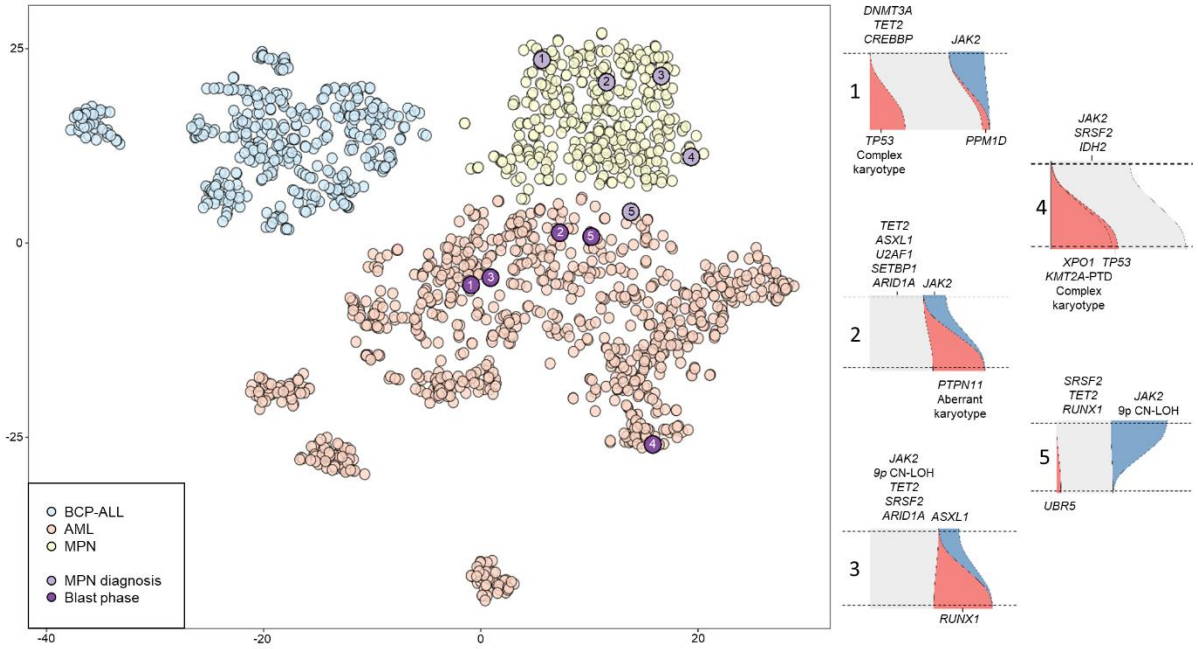

Figure S3: Changing transcriptomic profiles during progression to MPN-BP (cohort 4). The left side shows a UMAP based on the top 15% most variable genes between BCP-ALL (light blue), AML (light orange), and MPN-CP (light yellow) samples. Each dot represents a sample. The paired MPN-CP (light purple) 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.

2.6. Selected variants of MPN-CP and MPN-BP cases

Variants of patients in MPN-CP and MPN-BP were selected as indicated in Figure 6. Most frequently mutated genes as well as genes annotated to a specific pathway are illustrated. Information on chromosomal coordinates, aberration type, annotation and VAFs at both time points (MPN-CP and MPN-BP) are given.

Pat ID MPN-CP	Pat ID MPN-BP	*Gene*	Chromosomal position	Transcript ID	type	HGVSc	HGVSp	VAF [%] at MPN-CP	VAF [%] at MPN-BP
MLL_134389	MLL_15205	*ASXL1*	chr20:31022572	ENST00000375687	frameshift	c.2060_2061del	p.Cys687Tyrfs*30	30%	40%
MLL_134749	MLL_134751	*ASXL1*	chr20:31022978	ENST00000375687	frameshift	c.2464_2465insAA	p.Thr822Lysfs*3	9%	39%
MLL_134386	MLL_134395	*ASXL1*	chr20:31022286	ENST00000375687	nonsense	c.1772dup	p.Tyr591*	38%	48%
MLL_134392	MLL_134398	*ASXL1*	chr20:31022452	ENST00000375687	frameshift	c.1940dup	p.Gly649Trpfs*9	31%	42%
MLL_134750	MLL_134752	*ASXL1*	chr20:31022441	ENST00000375687	frameshift	c.1934dup	p.Gly646Trpfs*12	32%	35%
MLL_134382	MLL_134401	*ASXL1*	chr20:31021250	ENST00000375687	nonsense	c.1249C>T	p.Arg417*	49%	49%
MLL_25255	MLL_134147	*ASXL1*	chr20:31022263	ENST00000375687	nonsense	c.1748G>A	p.Trp583*	13%	0%
MLL_134390	MLL_134399	*BCOR*	chrX:39921394	ENST00000378444	nonsense	c.4426G>T	p.Glu1476*	0%	38%
MLL_134397	MLL_134753	*BCOR*	chrX:39923092	ENST00000378444	frameshift	c.3621dup	p.Gln1208Thrfs*8	0%	22%
MLL_134749	MLL_134751	*BHLHE22*	chr8:65494017	ENST00000321870	inframe_insertion	c.672_673insAGC	p.Gly224_Gly225insSer	0%	6%
MLL_10174	MLL_134391	*CBL*	chr11:119148991	ENST00000264033	missense	c.1211G>A	p.Cys404Tyr	32%	88%
MLL_134389	MLL_15205	*CEBPA*	chr19:33792422	ENST00000498907	missense	c.899G>T	p.Arg300Leu	0%	61%
MLL_134386	MLL_134395	*CEBPA*	chr19:33793258	ENST00000498907	frameshift	c.68dup	p.His24Alafs*84	0%	45%
MLL_134386	MLL_134395	*CEBPA*	chr19:33792387	ENST00000498907	nonsense	c.934C>T	p.Gln312*	0%	46%
MLL_134392	MLL_134398	*CREBBP*	chr16:3900638	ENST00000262367	missense	c.458C>T	p.Pro153Leu	42%	49%
MLL_134380	MLL_134384	*CTCF*	chr16:67650714	ENST00000264010	missense	c.1019A>G	p.His340Arg	37%	42%
MLL_134381	MLL_134388	*DNMT3A*	chr2:25467428	ENST00000264709	nonsense	c.1648G>T	p.Gly550*	48%	23%
MLL_21375	MLL_134393	*DNMT3A*	chr2:25457243	ENST00000264709	missense	c.2644C>T	p.Arg882Cys	39%	41%
MLL_10174	MLL_134391	*EFNA2*	chr19:1286196	ENST00000215368	inframe_insertion	c.29_40dup	p.Pro10_Leu13dup	0%	19%
MLL_134397	MLL_134753	*ETS2*	chr21:40178042	ENST00000360214	splice_site	c.-3_-2delinsGC	p.splice site mutation	30%	7%
MLL_134394	MLL_134400	*ETS2*	chr21:40178042	ENST00000360214	splice_site	c.-3_-2delinsGC	p.splice site mutation	20%	21%

MLL_25255	MLL_134147	*ETS2*	chr21:40178042	ENST00000360214	splice_site	c.-3_-2delinsGC	p.splice site mutation	26%	26%
MLL_19188	MLL_15202	*GATA2*	chr3:128204699	ENST00000341105	frameshift	c.741dup	p.Thr248Hisfs*34	0%	28%
MLL_134392	MLL_134398	*GATA4*	chr8:11566346	ENST00000642612	inframe_insertion	c.532_537dup	p.Ala178_Ala179dup	0%	19%
MLL_134380	MLL_134384	*GRIN2A*	chr16:9934614	ENST00000396573	missense	c.1541T>A	p.Ile514Asn	7%	0%
MLL_134382	MLL_134401	*IDH1*	chr2:209113112	ENST00000345146	missense	c.395G>A	p.Arg132His	45%	40%
MLL_134748	MLL_134385	*IDH2*	chr15:90631934	ENST00000330062	missense	c.419G>A	p.Arg140Gln	30%	41%
MLL_134747	MLL_12437	*IDH2*	chr15:90631934	ENST00000330062	missense	c.419G>A	p.Arg140Gln	41%	42%
MLL_134380	MLL_134384	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	39%	46%
MLL_134747	MLL_12437	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	41%	34%
MLL_134381	MLL_134388	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	27%	15%
MLL_19188	MLL_15202	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	49%	0%
MLL_134389	MLL_15205	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	75%	51%
MLL_134387	MLL_45460	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	20%	0%
MLL_134749	MLL_134751	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	18%	89%
MLL_134383	MLL_134396	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	58%	96%
MLL_134383	MLL_134396	*JAK2*	chr9:5054849	ENST00000381652	missense	c.901G>T	p.Gly301Trp	26%	0%
MLL_134392	MLL_134398	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	14%	0%
MLL_134750	MLL_134752	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	25%	12%
MLL_134390	MLL_134399	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	38%	0%
MLL_134397	MLL_134753	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	21%	59%
MLL_134394	MLL_134400	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	90%	97%
MLL_134382	MLL_134401	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	89%	97%
MLL_25255	MLL_134147	*JAK2*	chr9:5073770	ENST00000381652	missense	c.1849G>T	p.Val617Phe	78%	87%
MLL_134380	MLL_134384	*KRAS*	chr12:25398285	ENST00000256078	missense	c.34G>A	p.Gly12Ser	0%	4%
MLL_21375	MLL_134393	*LRP1B*	chr2:141093209	ENST00000389484	missense	c.12091G>T	p.Ala4031Ser	9%	0%
MLL_134392	MLL_134398	*LRP1B*	chr2:141473564	ENST00000389484	missense	c.6001A>G	p.Ile2001Val	60%	51%
MLL_21375	MLL_134393	*MESP1*	chr15:90294291	ENST00000300057	frameshift	c.171_172insAG	p.Pro58Serfs*13	29%	80%
MLL_21375	MLL_134393	*MESP1*	chr15:90294271	ENST00000300057	frameshift	c.191_192insG	p.Arg65Profs*7	17%	29%
MLL_134392	MLL_134398	*MESP1*	chr15:90294291	ENST00000300057	frameshift	c.171_172insAG	p.Pro58Serfs*13	82%	33%
MLL_134392	MLL_134398	*MESP1*	chr15:90294271	ENST00000300057	frameshift	c.191_192insG	p.Arg65Profs*7	33%	18%

MLL_134394	MLL_134400	*MESP1*	chr15:90294271	ENST00000300057	frameshift	c.191_192insG	p.Arg65Profs*7	26%	0%
MLL_134748	MLL_134385	*MPL*	chr1:43815009	ENST00000372470	missense	c.1544G>T	p.Trp515Leu	67%	3%
MLL_21375	MLL_134393	*MPL*	chr1:43815009	ENST00000372470	missense	c.1544G>T	p.Trp515Leu	91%	93%
MLL_21375	MLL_134393	*MPL*	chr1:43818306	ENST00000372470	missense	c.1771T>G	p.Tyr591Asp	95%	98%
MLL_134390	MLL_134399	*MYB*	chr6:135511292	ENST00000341911	missense	c.334C>T	p.Pro112Ser	0%	25%
MLL_134390	MLL_134399	*MYB*	chr6:135511397	ENST00000341911	missense	c.439T>C	p.Trp147Arg	0%	29%
MLL_134381	MLL_134388	*MYC*	chr8:128750864	ENST00000377970	missense	c.401C>G	p.Pro134Arg	47%	45%
MLL_134390	MLL_134399	*NRAS*	chr1:115258748	ENST00000369535	missense	c.34G>A	p.Gly12Ser	0%	17%
MLL_134382	MLL_134401	*NRAS*	chr1:115256521	ENST00000369535	missense	c.190T>G	p.Tyr64Asp	0%	10%
MLL_134747	MLL_12437	*PAK3*	chrX:110435396	ENST00000360648	missense	c.980G>T	p.Gly327Val	11%	0%
MLL_134749	MLL_134751	*PCBP2*	chr12:53849783	ENST00000359462	splice_site	c.243+2_244del	p.splice site mutation	18%	16%
MLL_134749	MLL_134751	*PCBP2*	chr12:53853185	ENST00000359462	splice_site	c.375+1_376-1del	p.splice site mutation	6%	5%
MLL_134749	MLL_134751	*PCBP2*	chr12:53854926	ENST00000359462	splice_site	c.504+1_505-1del	p.splice site mutation	15%	12%
MLL_134749	MLL_134751	*PCBP2*	chr12:53856349	ENST00000359462	splice_site	c.591+2_685del	p.splice site mutation	6%	0%
MLL_134392	MLL_134398	*PTPN11*	chr12:112888199	ENST00000351677	missense	c.215C>T	p.Ala72Val	0%	32%
MLL_134390	MLL_134399	*PTPN11*	chr12:112888189	ENST00000351677	missense	c.205G>A	p.Glu69Lys	0%	14%
MLL_19188	MLL_15202	*RUNX1*	chr21:36231831	ENST00000344691	nonsense	c.472C>T	p.Gln158*	38%	43%
MLL_134389	MLL_15205	*RUNX1*	chr21:36171728	ENST00000344691	nonsense	c.756G>A	p.Trp252*	40%	34%
MLL_134394	MLL_134400	*RUNX1*	chr21:36252880	ENST00000344691	missense	c.401T>C	p.Leu134Pro	44%	45%
MLL_21375	MLL_134393	*SETBP1*	chr18:42531907	ENST00000282030	missense	c.2602G>A	p.Asp868Asn	41%	44%
MLL_134392	MLL_134398	*SETBP1*	chr18:42531917	ENST00000282030	missense	c.2612T>C	p.Ile871Thr	32%	46%
MLL_134749	MLL_134751	*SETD1B*	chr12:122243933	ENST00000267197	missense	c.466G>A	p.Gly156Arg	28%	0%
MLL_134390	MLL_134399	*SF3B1*	chr2:198266834	ENST00000335508	missense	c.2098A>G	p.Lys700Glu	4%	30%
MLL_134747	MLL_12437	*SMARCA1*	chrX:128582408	ENST00000371122	missense	c.3043C>A	p.Arg1015Ser	11%	0%
MLL_134750	MLL_134752	*SPI1*	chr11:47377044	ENST00000378538	missense	c.547G>T	p.Gly183Cys	0%	13%
MLL_134748	MLL_134385	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>T	p.Pro95Leu	38%	45%
MLL_134747	MLL_12437	*SRSF2*	chr17:74732935	ENST00000392485	inframe_deletion	c.284_307del	p.Pro95_Arg102del	30%	23%
MLL_134381	MLL_134388	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>G	p.Pro95Arg	9%	7%
MLL_19188	MLL_15202	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>A	p.Pro95His	49%	43%
MLL_10174	MLL_134391	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>T	p.Pro95Leu	3%	0%

MLL_134389	MLL_15205	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>G	p.Pro95Arg	42%	35%
MLL_21375	MLL_134393	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>T	p.Pro95Leu	23%	46%
MLL_134383	MLL_134396	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>A	p.Pro95His	37%	46%
MLL_134750	MLL_134752	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>A	p.Pro95His	40%	39%
MLL_134397	MLL_134753	*SRSF2*	chr17:74732959	ENST00000392485	inframe_insertion	c.281_283dup	p.Arg94dup	36%	42%
MLL_134394	MLL_134400	*SRSF2*	chr17:74732935	ENST00000392485	inframe_deletion	c.284_307del	p.Pro95_Arg102del	20%	22%
MLL_134382	MLL_134401	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>A	p.Pro95His	46%	50%
MLL_25255	MLL_134147	*SRSF2*	chr17:74732959	ENST00000392485	missense	c.284C>A	p.Pro95His	47%	44%
MLL_134381	MLL_134388	*TET2*	chr4:106164778	ENST00000380013	nonsense	c.3646C>T	p.Arg1216*	25%	15%
MLL_19188	MLL_15202	*TET2*	chr4:106156747	ENST00000380013	nonsense	c.1648C>T	p.Arg550*	47%	40%
MLL_19188	MLL_15202	*TET2*	chr4:106155829	ENST00000380013	nonsense	c.730A>T	p.Lys244*	51%	40%
MLL_134389	MLL_15205	*TET2*	chr4:106155599	ENST00000380013	nonsense	c.500C>G	p.Ser167*	32%	37%
MLL_134387	MLL_45460	*TET2*	chr4:106157240	ENST00000380013	nonsense	c.2141C>G	p.Ser714*	37%	36%
MLL_21375	MLL_134393	*TET2*	chr4:106156239	ENST00000380013	frameshift	c.1142dup	p.Lys382Glnfs*8	45%	49%
MLL_134383	MLL_134396	*TET2*	chr4:106180796	ENST00000380013	missense	c.3824G>A	p.Gly1275Glu	44%	39%
MLL_134392	MLL_134398	*TET2*	chr4:106158346	ENST00000380013	nonsense	c.3247C>T	p.Gln1083*	35%	41%
MLL_134397	MLL_134753	*TET2*	chr4:106164916	ENST00000380013	missense	c.3784C>T	p.Arg1262Trp	42%	47%
MLL_25255	MLL_134147	*TET2*	chr4:106158219	ENST00000380013	frameshift	c.3121_3124del	p.Phe1041Thrfs*13	43%	36%
MLL_25255	MLL_134147	*TET2*	chr4:106197149	ENST00000380013	nonsense	c.5482C>T	p.Gln1828*	10%	0%
MLL_134380	MLL_134384	*TP53*	chr17:7578394	ENST00000269305	missense	c.536A>G	p.His179Arg	1%	39%
MLL_134747	MLL_12437	*TP53*	chr17:7579355	ENST00000269305	missense	c.332T>G	p.Leu111Arg	0%	6%
MLL_134387	MLL_45460	*TP53*	chr17:7578556	ENST00000269305	splice_site	c.376-2A>T	p.splice site mutation	0%	32%
MLL_134382	MLL_134401	*TP53*	chr17:7577545	ENST00000269305	missense	c.736A>G	p.Met246Val	0%	14%
MLL_134392	MLL_134398	*U2AF1*	chr21:44514777	ENST00000291552	missense	c.470A>G	p.Gln157Arg	37%	38%
MLL_10174	MLL_134391	*ZRSR2*	chrX:15809102	ENST00000307771	frameshift	c.88del	p.Arg30Valfs*8	46%	87%
MLL_10174	MLL_134391	*ZRSR2*	chrX:15809106	ENST00000307771	missense	c.91C>T	p.Arg31Trp	26%	0%

3. References

1. Delic S, Rose D, Kern W, Nadarajah N, Haferlach C, Haferlach T, Meggendorfer M. 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 7/22/2016.

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)**:** 2090-2090.

3. Johannessen B, Sveen A, Skotheim RI. TIN: An R Package for Transcriptome Instability Analysis. *Cancer Inform* 2015; **14:** 109-112.

4. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A*, et al.* Proteomics. Tissue-based map of the human proteome. *Science* 2015 Jan 23; **347**(6220)**:** 1260419.

5. Singh Nanda J, Kumar R, Raghava GP. dbEM: A database of epigenetic modifiers curated from cancerous and normal genomes. *Sci Rep* 2016 Jan 18; **6:** 19340.

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

7. 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 8/15/2013; **29**(16)**:** 2041-2043.

8. 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 8/2018; **15**(8)**:** 591-594.

9. 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 Apr 15; **32**(8)**:** 1220-1222.

10. 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 May; **581**(7809)**:** 434-443.

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

12. 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.

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

14. 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 Mar; **31**(3)**:** 448-460.

15. 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 Oct 21; **20**(1)**:** 213.

16. 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 May; **28**(5)**:** 511-515.

17. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010 Jan 1; **26**(1)**:** 139-140.