Supplemental Material

Supplementary Text S1. Statistical analysis. Gene expression data were gcrma-normalized (Wu et al. 2004). To assess presence and absence of gene expression independently of Affymetrix-mismatch-probesets, the "Presence-Absence calls with Negative Probe sets (PANP)" algorithm (Warren et al. 2007) was used. Differential gene expression was assessed using empirical Bayes statistics in linear models for microarray data (Smyth 2004). *P*-values were adjusted for multiple testing controlling the false-discovery-rate as defined by Benjamini and Hochberg (Benjamini and Hochberg 1995). Expression profiles of 434 samples (233 MMCs, 14 BMPCs, 12 PPCs, 12 MGUS, 40 HMCLs (the same 20 HMCLs on different microarrays in Heidelberg/Montpellier-group 1 (HM1) and 2 (HM2)), 13 MBCs, 64 WBM, 19 MSCs, 5 $CD3^+$, 5 $CD14^+$, 5 $CD15^+$, 5 $CD34^+$ and 7 OCs) divided in Heidelberg/Montpellier-group 1 (HM1; n=113, MM n=65) and HM2 (n=257, MM n=168) were analyzed. Event-free and overall survival (Goldschmidt et al. 2003) were investigated for the 168 patients (48 HM1, 120 HM2) undergoing high-dose chemotherapy and autologous stem cell transplantation using Cox's proportional hazard model. First, BMP6-expression was taken as a continuous variable. Secondly, BMP6-expression was tested in a Cox-model together with either B2M or ISS. Next, two groups of patients with high ($BMP6^{high}$, greater or equal the median) and low ($BMP6^{low}$, below the median) BMP6-expression were delineated. Findings were validated using the same strategy on the independent group of 345 patients from the Little Rock-group. qRT-PCR data were analyzed in analogy to previously published protocols (Mahtouk et al. 2005).

For myeloma cells, association of chromosomal aberrations and clinical parameters with gene expression was assessed using two-sample t-statistic. Differences in clinical parameters between defined groups were investigated by analysis-of-variance. Correlation was measured using the

Spearman and Pearson correlation coefficient. Correlation with categorical variables was measured using the Kendall's tau coefficient.

An effect was considered as statistically significant if the *P*-value of its corresponding statistical test was not greater than 5 %. Statistical computations were performed using R (R Development Core Team 2008) version 2.5.1 and Bioconductor (Gentleman et al. 2004), version 2.0.

Supplementary Text S2. Calculation of the gene expression based proliferation index. The gene expression based proliferation index is calculated as follows. In brief, genes are selected based on genes over-expressed in proliferating cells (malignant: human myeloma cell lines (HMCLs), benign: polyclonal plasmablastic cells (PPCs)) compared to non-proliferating cells (normal bone marrow plasma cells (BMPCs) and memory B-cells (MBCs)). Here, four comparisons between the groups are made (i) HMCL vs. MBC, (ii) HMCL vs. BMPC, (iii) PPC vs. BMPC and (iv) PPC vs. MBC) by a one-sided t-test, with the alternative hypothesis being that expression values of HMCLs and PPCs values are greater compared to BMPCs and MBCs in each comparison. *P*-values are permutation-adjusted regarding a family wise error rate with an α-level of 0.025. To adjust for comparing each group twice, the α-level is halved to 0.0125. Only genes being statistically significant in each of the 4 comparisons are retained for the index. To select biologically (in terms of proliferation) relevant genes, only genes matching with the gene-ontology term „cell proliferation" or „cell cycle" were retained. Thus, 50 genes (57 probesets) represent the final index. For genes with more than one probeset per gene, the one with the highest variance within the training group is selected. The index is calculated as follows: As proliferation-genes determined as stated above are over-expressed by definition, the individual gene expression based proliferation index for each sample is calculated as the sum of expression values of each of the 50 genes in an individual sample. For genes not expressed as judged by PANP, the expression level of the respective gene is defined as 0.

Reference List

Benjamini Y and Hochberg Y. (1995). *Journal of the Royal Statistical Society Series B,* 57, 289-300.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY and Zhang J. (2004). *Genome Biol,* 5, R80.

Goldschmidt H, Sonneveld P, Cremer FW, van der HB, Westveer P, Breitkreutz I, Benner A, Glasmacher A, Schmidt-Wolf IG, Martin H, Hoelzer D, Ho AD and Lokhorst HM. (2003). *Ann Hematol,* 82, 654-659.

Mahtouk K, Hose D, Reme T, De Vos J, Jourdan M, Moreaux J, Fiol G, Raab M, Jourdan E, Grau V, Moos M, Goldschmidt H, Baudard M, Rossi JF, Cremer FW and Klein B. (2005). *Oncogene,* 24, 3512-3524.

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing: Vienna, Austria.

Smyth GK. (2004). *Stat Appl Genet Mol Biol,* 3, Article3.

Warren, P., Taylor, D., Martini, P. G. V., JAckson, J., and Bienkowska, J. PANP - a New Method of Gene Detection on Oligonucleotide Expression Arrays. Bioinformatics and Bioengineering, 2007.BIBE 2007.Proceedings of the 7th IEEE International Conference on 14-17 Oct.2007 , 108-115. 2007.

Wu Z, Irizarry RA, Gentleman RC, Martinez-Murillo F and Spencer F. (2004). *Journal of the American Statistical Association,* 99, 909-917.