Genes with a spike expression are clustered in chromosome (sub)bands and spike (sub)bands have a powerful prognostic value in patients with multiple myeloma

Alboukadel Kassambara,[1] Dirk Hose,[2] Jérôme Moreaux,[1,3] Brian A. Walker,[4] Alexei Protopopov,[6] Thierry Reme,[1,3] Franck Pellestor,[1,3,5] Véronique Pantesco,[1] Anna Jauch,[2] Gareth Morgan,[4] Hartmut Goldschmidt,[2] and Bernard Klein[1,3,5]

[1]INSERM U1040, Montpellier, France; [2]Medizinische Klinik V, Universitätsklinikum Heidelberg and Nationales Centrum für Tumorerkrankungen, Heidelberg, Germany; [3]CHU Montpellier, Institute of Research in Biotherapy, Montpellier, France; [4]Section of Haemato-Oncology, The Institute of Cancer Research, London, United Kingdom; [5]Université Montpellier 1, UFR Médecine, Montpellier, France, and [6]Jerome Lipper Multiple Myeloma Center, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

ABSTRACT

Background
Genetic abnormalities are common in patients with multiple myeloma, and may deregulate gene products involved in tumor survival, proliferation, metabolism and drug resistance. In particular, translocations may result in a high expression of targeted genes (termed spike expression) in tumor cells. We identified spike genes in multiple myeloma cells of patients with newly-diagnosed myeloma and investigated their prognostic value.

Design and Methods
Genes with a spike expression in multiple myeloma cells were picked up using box plot probe set signal distribution and two selection filters.

Results
In a cohort of 206 newly diagnosed patients with multiple myeloma, 2587 genes/expressed sequence tags with a spike expression were identified. Some spike genes were associated with some transcription factors such as MAF or MMSET and with known recurrent translocations as expected. Spike genes were not associated with increased DNA copy number and for a majority of them, involved unknown mechanisms. Of spiked genes, 36.7% clustered significantly in 149 out of 862 documented chromosome (sub)bands, of which 53 had prognostic value (35 bad, 18 good). Their prognostic value was summarized with a spike band score that delineated 23.8% of patients with a poor median overall survival (27.4 months *versus* not reached, $P<0.001$) using the training cohort of 206 patients. The spike band score was independent of other gene expression profiling-based risk scores, t(4;14), or del17p in an independent validation cohort of 345 patients.

Conclusions
We present a new approach to identify spike genes and their relationship to patients' survival.

Key words: spike (sub)bands, myeloma, molecular biology, survival, prognosis.

Citation: Kassambara A, Hose D, Moreaux J, Walker BA, Protopopov A, Reme T, Pellestor F, Pantesco V, Jauch A, Morgan G, Goldschmidt H, and Klein B. Genes with a spike expression are clustered in chromosome (sub)bands and spike (sub)bands have a powerful prognostic value in patients with multiple myeloma. Haematologica 2012;97(4):622-630. doi:10.3324/haematol.2011.046821

©2012 Ferrata Storti Foundation. This is an open-access paper.

Acknowledgments: we acknowledge the support of the Biological Research Centre at the Royal Marsden Hospital.

Funding: this work was supported by grants from the Ligue Nationale Contre le Cancer (équipe labellisée), Paris, France, from ARC (France), EU FP7 (OVERMYR), the Hopp-Foundation, Germany, the University of Heidelberg, Germany, the National Centre for Tumor Diseases, Heidelberg, Germany, the Tumorzentrum Heidelberg/Mannheim, Germany, the Deutsche Krebshilfe, Bonn, Germany, and the Deutsche Forschungsgemeinschaft, Bonn, Germany. We thank the Microarray Core Facility of IRB (http://irb.montp.inserm.fr/en/index.php?page=Plateau&IdEquipe=6) and the cytometry platform of the Institute of Research in Biotherapy (http://irb.montp.inserm.fr/en/index.php?page=Plateau&IdEquipe=3, Montpellier Rio Imaging).

Manuscript received on May 1, 2011. Revised version arrived on November 7, 2011. Manuscript accepted on November 9, 2011.

Correspondence:
Bernard Klein, Institut de Recherches en Biothérapie, CHU Montpellier, Hopital St Eloi, Av Augustin Fliche 34285 MONTPELLIER Cedex 5, France.
Phone: international +33.4.67330190.
Fax: international +33.4.67337905.
E-mail: bernard.klein@inserm.fr

The online version of this article has a Supplementary Appendix.

Introduction

Multiple myeloma is a molecularly heterogeneous disease with recurrent gene translocations, amplifications and deletions in the MM cells detected by conventional cytogenetics, interphase fluorescence *in situ* hybridization (iFISH) and more recently by whole genome DNA copy number microarray analysis.[1,2] MM is broadly divided into hyperdiploid MM (45% of cases) and non-hyperdiploid MM (40%). Hyperdiploid MM harbor odd chromosome duplications and non-hyperdiploid MM recurrent translocations. These recurrent translocations are t(11;14)(q13;q32) involving *CCND1* gene and t(4;14)(p16.3;q32) involving *MMSET/WHSC1* and *FGFR3* genes in 15% of patients each, and less frequent translocations involving *C-MAF*, *MAFB* and *cyclin D3*. t(4;14)(p16.3;q32) translocation is associated with a poor prognosis.[3] Additional genetic events involve chromosome 13 deletion in 50% of the patients, chromosome 17p deletion in 10%, chromosome 12 deletion in 12%, chromosome 1p loss in 30% and 1q gain in 36%.[1,2,4] Point mutations have also been identified targeting the ras pathway[5,6] and the NK-κB pathway.[7,8] Array comparative genomic hybridization (CGH) or single nucleotide polymorphism (SNP) microarrays have made it possible to confirm the classification of patients into two major groups, hyperdiploid MM characterized by odd chromosome duplications and non-hyperdiploid MM, and have identified numerous microalterations, some with prognostic value.[2,3,9,10]

These translocations, amplifications or deletions could result in either gain- or loss-of-function in genes encoding for proteins that may control tumor cell survival, proliferation and motility.[11] They could result in aberrant- or overexpression of genes directly targeted by the genetic abnormality or of genes whose expression is induced by the product of the targeted gene.[12,13] This is the case since whole genome transcriptome analysis of purified MM cells from large cohort of patients[14-16] made it possible to classify patients into seven molecular subgroups, five of which associated with recurrent translocations or hyperdiploidy. Subsets of genes whose expression in MM cells is associated with poor or good prognosis were also indentified using whole genome transcriptome microarray.[17-21]

In the case of recurrent translocations, the translocated genes have a spike expression, *i.e.* they are highly expressed in MM cells from the fraction of patients with the translocation compared to the remaining patients of the whole cohort. The most prominent examples are *FGFR3* and *MMSET* in patients with t(4;14)(p16.3;q32), and *CCND1* in those with t(11;14)(q13;q32).[22-24] Given these observations, we hypothesized that a gene with strong overexpression in MM cells of a fraction of patients (termed spike gene) could have some implications in MM disease. We present a method to identify and rank spike genes in patients with newly-diagnosed MM. We identified 2587 genes/expressed sequence tags (EST) with a spike expression starting from a cohort of 206 newly-diagnosed patients. Genes involved in known recurrent translocations in MM had the highest spike score. Spike genes were significantly enriched in 149 out of the 862 identified chromosome bands, and 53 of these 149 spike bands were associated with either bad or good prognosis making it possible to define a spike band score with high prognostic value.

Design and Methods

Patients' samples and gene expression data

MM cells were purified from the 206 patients with newly-diagnosed MM after written informed consent was given at the University Hospitals of Heidelberg (Germany) or Montpellier (France) as described elsewhere.[25] The clinical characteristics of the Heidelberg and Montpellier (HM) cohort are provided in *Online Supplementary Table S1*. The study was approved by the ethics boards of the University Hospitals of Heidelberg and Montpellier. Gene expression of purified MM cells was profiled using Affymetrix U133 2.0 plus microarrays as described elsewhere[26] and data were normalized using the MAS5 Affymetrix algorithm with a scaling factor of 100. The .CEL and MAS5 files are deposited in the ArrayExpress public database (*http://www.ebi.ac.uk/arrayexpress/*), under accession number E-MTAB-362, user: Reviewer_E-MTAB-362, password: 1284054153439. We also used publicly available MAS5 normalized GEP data (GEO, *http://www.ncbi.nlm.nih.gov/geo/*, accession number GSE2658) from purified MM cells of a cohort of 345 patients (UAMS-TT2 cohort) treated with the Total Therapy 2 protocol at the University of Arkansas for Medical Sciences (UAMS, Little Rock, USA).[27] As iFISH data were not available for UAMS-TT2 patients, t(4;14) translocation was evaluated using *MMSET* spike expression[21] and del17p13 surrogated by the level of *TP53*.[28] Publicly-available gene expression profiling (GEP) data from 39 human myeloma cell lines (HMCL) were used (*http://www.ebi.ac.uk/arrayexpress/*, accession numbers E-TABM-937 and E-TABM-1088). The phenotypic, functional and molecular characteristics of these 39 HMCL have been reported elsewhere.[26]

Identification of spike genes and definition of a spike score

Spike genes are defined as genes whose expression in MM cells is high in a subgroup of patients only. We used the following series of filters to pick up these spike probe sets. The first filter used the coefficient of variation (CV), *i.e.* the ratio of standard deviation to mean, and the maximum signal (MS) of probe set signals across samples. The CV and the MS of each of the 54613 probe sets on Affymetrix U133 2.0 plus arrays were computed and their distributions are shown in *Online Supplementary Figure S1A*. There is a large number of probe sets with low variability and/or low expression. Out of the 54613 probe sets, 9151 had a CV or MS higher than the median values of the CV and MS in the HM cohort (*Online Supplementary Figure S1A*). The second filter used the box plot probe set signal distribution to identify extreme values ("outliers"). It selected probe sets whose signal in some patients was ≥ Q3+3*IQR, Q3 being quartile 3 of the probe set signals among patients' samples and IQR the interquartile range (Q3-Q1). The number of probe sets for which the percentage of patients with a signal ≥ Q3+3*IQR was higher or equal to a given percentage is plotted in *Online Supplementary Figure S1B*. For about half the 9151 filter 1 probe sets, the frequency was ≥1%. We selected the 4223 probe sets with a frequency >1%, probing for 2587 unique genes/EST with known chromosomal localization (*Online Supplementary Figure S1B*). Once spike probe sets had been selected, they were ranked using a score measuring the distance of the mean expression of signals in patients with spike expression (*i.e.* ≥ Q3+3*IQR) from the mean value of signals in patients with non-spike expression. The following score was used: $\log_2[(\text{Mean(signals} \geq Q3+3*IQR) - \text{Mean(signals} < Q3+3*IQR)/SD(\text{signals} < Q3+3*IQR)]$. The higher the spike score of a given probe set is, the more likely a spike expression is. For a given gene/EST, when several spike probe sets were identified, the probe set with the highest spike score was used.

Spike gene correlation network analysis

Correlation network analysis was done for two of the highest spike genes: *MMSET* and *MAF*. Firstly the top 100 spike genes significantly correlated with *MMSET* or *MAF* were selected using Pearson's correlation test adjusted by Benjamini-Hochberg multiple testing correction ($P \leq 0.05$). The correlation matrix of these top 100 correlated spike genes was then calculated and visualized using Cytoscape 2.6.3 software[29] considering only significant correlation links with Pearson's correlation coefficient ≥ 0.5.

Copy number analysis using Affymetrix genechip human mapping 250K Nsp array

Independent data from an Institute of Cancer Research (ICR, London) cohort,[2] comprising a MM expression array (n=258) and MM 250K Nsp mapping array (n=114), were used to further investigate the link between spiking and copy number alteration. This data set, available under GEO accession number GSE21349, also includes a 250K Nsp mapping array of peripheral white blood cells (n = 80) used as the normal counterpart.

Quality control, probe level normalization and copy number analysis were carried out with Affymetrix Genotyping Console software (GTC) 4.0. Copy number analysis was performed using peripheral white blood cells for comparison and the default settings in GTC. For copy number segment analysis, the GTC segment-reporting tool was used to identify regions in the genome with a setting of at least 40 markers showing consensus for gain or loss spanning at least a 100-kb region. Considering that the median spacing of the markers of this array is 2.5 Kb, one would expect 40 SNP markers per 100-kb region. Results of the segment analysis were used to compare aberrant regions.

Fluorescence in situ hybridization

iFISH analysis was performed on CD138-purified plasma cells as described previously,[19,30,31] using a set of probes for the chromosomal regions 1q21, 6q21, 8p21, 9q34, 11q23, 11q13, 13q14.3, 14q32, 15q22, 17p13, 19q13, and 22q11, as well as the translocations t(4;14)(p16.3;q32.3) and t(11;14)(q13;q32.3) (Poseidon-probes, Madison, WI, USA). Ploidy status and clonal/subclonal aberrations for a single aberration (*i.e.* present in ≥ 60% *versus* 20-59% of assessed MM cells) were defined as published.[32] A modified copy number score (CS)[32,33] (excluding gains of 1q21) and the score described by Wuilleme *et al.*(CSW),[34] using chromosomes 5, 15, and 19, were applied to assess ploidy. As regards the absolute number of chromosomal aberrations, 163 patients were assessed for the t(4;14)(p16.3;q32.3) and t(11;14)(q13;q32,3) translocations as well as numerical aberrations of the chromosomal regions 11q13, 11q23, 1q21, 17p13, 13q14.3, and 14q32.

Spectral karyotyping on human myeloma cell lines

Spectral karyotyping (SKY) was performed on 12 HMCL by Dr. Protopopov (DFCI, Boston, USA). The image acquisition and identification of breakpoints on the SKY-painted chromosomes was performed as described previously.[35,36]

Prognostic value of chromosome (sub)bands containing spiked genes

Among the 862 reported chromosome bands and sub-bands[37] (*http://genome.ucsc.edu/*), the "spike" (sub)bands with significantly increased frequency of spike genes out of the genes located in the (sub)band were identified using a χ^2 test and Benjamini Hochberg multiple testing correction ($P \leq 0.05$). A minimal number of four spike genes/(sub)band was chosen to make it possible to perform a χ^2 test with calculated numbers of four or more. For each patient and for each spike (sub)band, the mean expression of the signals of spike genes located in this band was computed and termed spike (sub)band expression (SBE). The prognostic value of the SBE of each band, computed using a maximally selected rank test from R package MaxStat (http://cran.r-project.org/web/packages/maxstat/index.html) on the HM patient cohort and the Benjamini Hochberg multiple correction, yielded 55 spike (sub)bands with prognostic value.

Building a spike band score

A spike band score (SBS) was built for each patient to group the information of the 53 prognostic (sub)bands within one parameter. For each prognostic (sub)band and for each patient, the product of the Wald statistic value and odd's ratio determined with the MaxStat package was weighted by +1 if the patient's SBE was above the Maxstat cutoff, and by –1 if it was below or equal to this cutoff. The SBS of a given patient is the sum of these parameters for the 53 prognostic (sub)bands. Patients from the same cohort were ranked according to increasing SBS and for a given value S, the difference in survival of patients with a SBS < S or ≥ S was computed, making it possible to define the SBS value with a maximum difference in survival using the maximally selected rank test from R package MaxStat.

Statistical analysis

Gene expression profiles were analyzed with our bioinformatics platform (RAGE, *http://rage.montp.inserm.fr*)[38] and with the Amazonia website (*http://amazonia.montp.inserm.fr/*).[39] The statistical significance of differences in survival between groups of patients was calculated using the log-rank test. An event was defined as relapse (for event-free survival) or as death (for overall survival). Multivariate analysis was performed using the Cox proportional hazards model. Survival curves were plotted using the Kaplan-Meier method. All these analyses were done with R.2.10.1 (*http://www.r-project.org/*) and bioconductor version 2.5.[40,41] Gene annotation and networks were generated through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, *www.ingenuity.com*).

Results

Classification of spike probe sets

We identified 2587 unique spike genes/EST (*Online Supplementary Table S2*) using the adjustable filter parameters described in the *Design and Methods* section (*Online Supplementary Figure S1*). We chose to start with a large spike gene list to apply further selection criteria based on patients' overall survival data to identify prognostic spike genes. The spike score of the 2587 genes/EST ranged from 1.85 to 10.08 according to the frequency distribution (Figure 1). To provide a quick overview of the value of the score of spike genes in the following figures, we arbitrarily split the spike genes/EST into five nearly equally sized categories (I to V according to decreasing spike score), each comprising 502-525 genes/EST (Figure 1 and *Online Supplementary Figure S2*). Affymetrix signals of the selected genes from each of the five categories are depicted in *Online Supplementary Figure S3*.

Spiking mechanisms

As reviewed recently,[12,13] gene overexpression in cancer cells could be explained by various mechanisms: translocation, amplification, point mutation, epigenetic control, or secondary activation by an amplified activator, or functional activation as for genes coding for proliferation-regulating proteins. Translocation acts as a spiking mechanism as evi-

denced by genes involved in known recurrent translocations in MM. These were among the genes with the highest spike score: 10.08 for *FGFR3* and 4.58 for *MMSET/WHSC1* involved in t(4;14)(p16.3;q32.3), 4.59 for *CCND1* involved in t(11;14)(q13;q32.3), and 4.58 for *MAF* involved in t(14;16)(q32.3;q23) (Figure 1). This was confirmed correlating iFISH data and expression of the genes involved in the translocation since spike *MMSET/WHSC1* expression was found in 96% of the patients with documented t(4;14)(p16.3;q32.3) by iFISH but in only 1% of patients with no t(4;14)(p16.3;q32.3); likewise, spike *CCND1* expression was found in 88% *versus* 2% of patients respectively with and without iFISH-documented t(11;14)(q13;q32.3) ($P \leq 0.05$, *Online Supplementary Table S3*).

A link between highly spiked genes and genomic rearrangements was further investigated using 12 HMCL

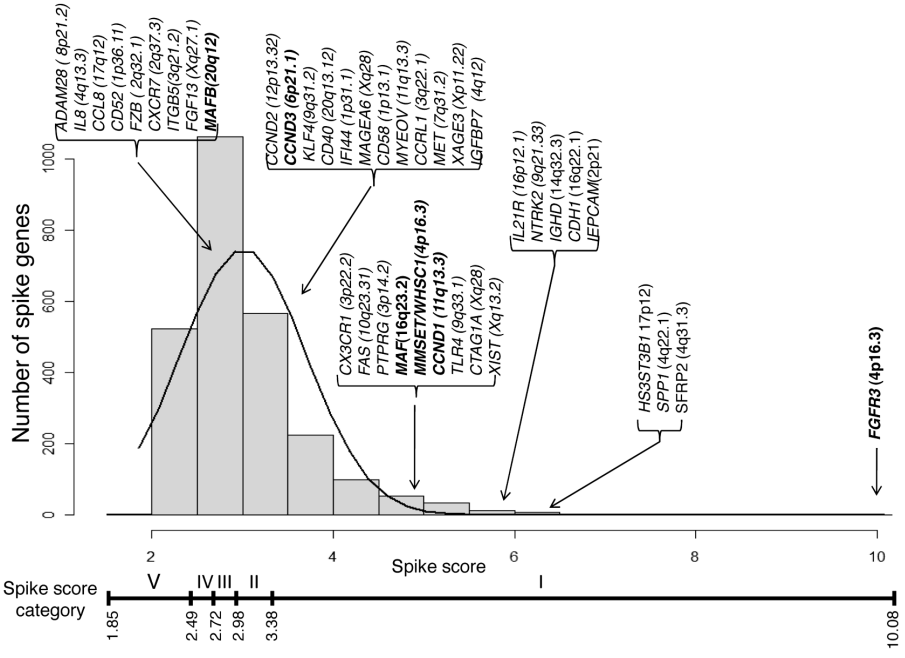
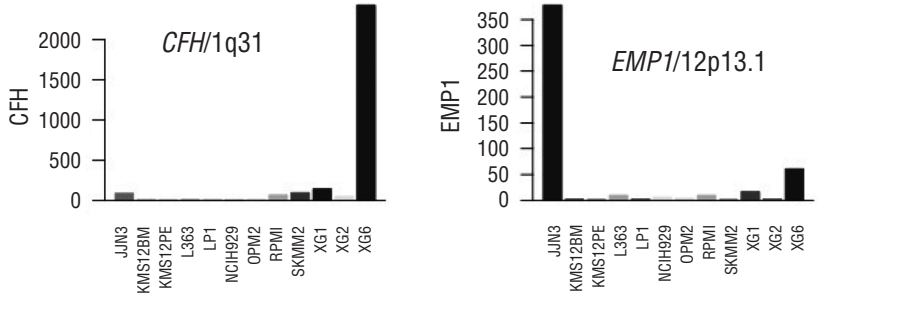


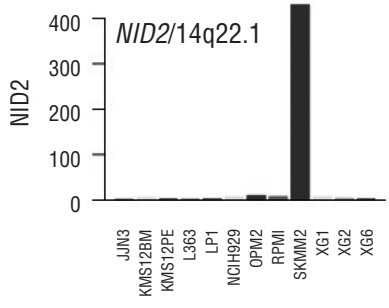
Figure 1. Distribution of spike genes according to their spike score. Genes known to be involved in MM-specific translocations are shown in bold. The spike score is a normalized measure of the distance of the mean expression of signals in patients with spike expression [i.e. ≥ Q3+3*(Q3-Q1)] from the mean expression of signals in patients with non-spike expression. The higher the score, the more spiked the probe set is. Spike probe sets were grouped in five nearly equally sized categories (I to V according to decreasing spike scores) comprising 502-525 probe sets. The most speaked genes are displayed.



XG6: der(1)t(1;7)(7qter-7q22::1q31-1q10::1q10-1qter)

JJN3: der(12)t(12;19)(19qter-19q13::12p13-12qter)

Figure 2. Link between spike expression and translocation. We investigated whether genes located in a given translocation-associated breakpoint could have a spike expression in the human myeloma cell lines presenting the given breakpoint only. Representative data are shown for *CFH* (1q31), *EMP1* (12p13), and *NID2* (14q22.1) genes. Barplots show the Affymetrix expression signal.



SKMM2: der(4)t(4;9;13;14) (:14p13-4q13::14q22-14q31::4q13-4q35::9q21-9q34::13q22-13qter

for which spectral karyotyping data were available. We investigated whether genes located in a cytoband involved in a translocation-associated breakpoint could have a spike expression in the HMCL with the given breakpoint. Between 3 and 19 translocation-associated breakpoints were identified in the individual HMCL, for a total of such breakpoints. One hundred and ninety-nine of the 521 highly spiked genes (category I) had a spike expression in these 12 HMCL and 25 of these 199 genes were spiked in the cell lines with a translocation-associated cytoband containing the gene locus (*Online Supplementary Table S4*). Three of the 25 breakpoint-associated spike genes have already been identified as caused by a translocation involving the *IGH* locus and *CCND1*, *MAF* and *MMSET*. For the remaining 22 genes, a translocation bringing the gene close to an active promoter could be one of the mechanisms explaining the spiking. The list of the 22 translocation-associated spike genes is provided in *Online Supplementary Table S4*. Of note, no novel translocations involving *IGH* or *IGL* loci were found. Figure 2 shows the data for the *CFH* gene located on 1q31 with a spike expression in XG6 HMCL only, presenting a der(1)t(1;7)(7qter-7q32::1q31-1q10::1q10-1qter). None of the 11 remaining HMCL had a rearrangement involving the 1q31 locus. Other examples are the *EMP1* (12p13.1) and *NID2* genes (14q22.1) with spike expression in the JJN3 and SKMM2 cell lines, respsectively (Figure 2). JJN3 has a der(12)t(12;19)(19qter-19q13::12p13-12qter) and is the only HMCL to show a rearrangement involving 12p13.1. SKMM2 has a der(4)t(4;9;13;14)(:4p13-4q13::14q22-14q31::4q13-4q35::9q21-9q34::13q22-13qter). The spike expression of the other genes is provided in *Online Supplementary File S1*.

Copy number alterations did not result in spiked gene expression

iFISH and GEP data were available for 163 of the 206 patients of the HM cohort. Chromosome regions 11q13, 9q34 and 15q22 were gained in 82 of these 163 patients without spike gene expression (*data not shown*). This is particularly the case for aberrant but low *CCND1* expression due to a gain in 11q13 copy number without t(11;14)(q13;q32.3) translocation (assayed with a CCDN1 iFISH probe on MM cells from 163 patients).[42] Only one of these 73 patients (1.4%) with 11q13 gain had a spike *CCND1* expression compared to 88% of the 25 patients with t(11;14)(q13;q32.3) translocation. A link between copy number alterations (CNA) and spike genes was also investigated in the ICR cohort of 81 patients for whom both Affymetrix U133 plus 2.0 transcriptome data and GeneChip Mapping 500K Array set data were available (GEO accession number GSE21349). Spiking was not significantly associated with CNA of small DNA regions. Using the ICR cohort, we found 687 spike genes located in the 269 DNA cytobands amplified in at least five patients (to be able to perform statistical comparisons). For all except one of these 687 spike genes, the spiking frequency was not significantly different between patients with CNA in the corresponding DNA region and those without (*data not shown*). Only the spiking frequency of *FBXO32* gene was significantly associated with increased 8q24.13 band.

To investigate whether genes could also show a spiked expression due to genetic events targeting their regulation (e.g. transcription factors targeted by translocations), we analyzed the network of genes correlated with the two most highly spiked genes coding for transcription factors: *MAF* and *MMSET*. As seen in Figure 3, nine spike genes were strongly and directly correlated with *MAF* expression. Five of these nine genes belong to category I spike genes, three to category II and one gene to category III. Eight spike genes were strongly correlated with *MMSET* expression, including five genes in category I, two in category II and one in category III. Thus, spike genes in categories I and II comprise genes involved in known translocations or genes activated by an amplified transcription factor. Recently, a census of human cancer genes was made[12,13] and an updated list of 427 causal cancer genes, about 2% of the human genes, is available at the Sanger database (*http://www.sanger.ac.uk/genetics/CGP/Census/*). Of interest, 18% (76 genes) of these 427 cancer genes were spiked in MM cells, indicating that the current method makes it possible to enrich census cancer genes significantly ($P<0.01$) (*Online Supplementary Table S5*).

The frequency of spike genes is significantly increased in selected chromosome (sub)bands

Spike genes were located on all 24 chromosomes. The frequencies of spike genes in the long arms of chromosomes was higher than those in short arms, mirroring the higher number of genes in long arms than in short arms (χ^2 test, $P< 10^{-4}$, *Online Supplementary Figure S4*). The five chromosomes with the highest frequencies of spike genes in their short arms were 18p (23%), 10p (17.4%), 12p (16.3%),

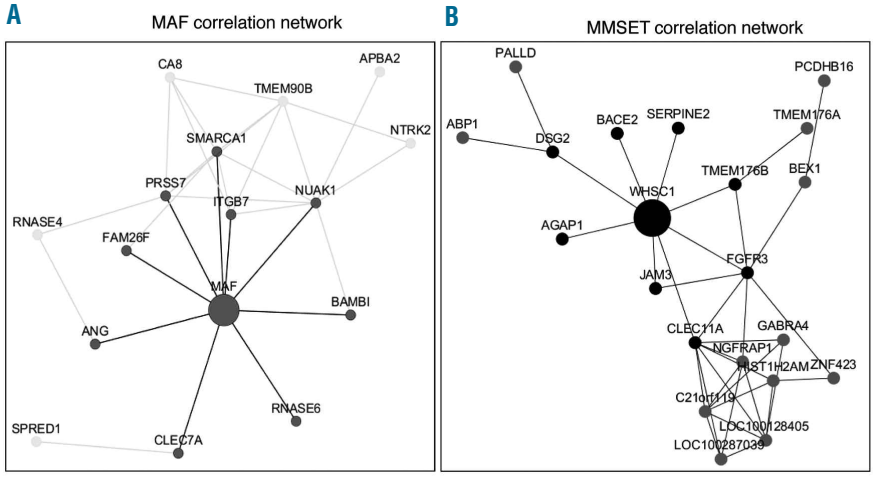


Figure 3. Networks of spike genes correlated with *MAF* or *MMSET* gene expression. The correlation network analysis was determined with two of the most spiked genes encoding for transcription factors (A) *MAF* and (B) *MMSET*. The top 100 spike genes significantly correlated with *MMSET* or *MAF* were identified using Pearson's correlation and multiple testing correction ($P \leq 0.05$). The correlation matrix of these top 100 correlated spike genes was then visualized using Cytoscape 2.6.3 software[29] considering genes with a Pearson's correlation coefficient ≥0.5. Data show the nodes of the most strongly correlated genes.

7p (15.3%) and 5p (15%), the five with the highest frequencies in their long arms were 4q (17.57%), 1q (17.2%), 6q (16.9%), 9q (15.6%) and 21q (15.5%). Using the HM cohort of patients, the frequency of spike genes in a given (sub)band was increased significantly ($P \leq 0.05$) in 149 of the 862 documented human chromosome (sub)bands (*http://genome.ucsc.edu/*) using Benjamini Hochberg multiple testing correction (*Online Supplementary Table S6*). These 149 (sub)bands are termed spike (sub)bands and their localization is shown in Figure 4A. Fifteen out of the 149 spike (sub)bands had more than ten spike genes. They are located on ten chromosomes (1, 4, 6, 9, 11, 12, 14, 15, 19 and X) (*Online Supplementary Table S7*). Figure 4B shows the three (sub)bands with ≥ ten spike genes for the highly spiked chromosome 1. The 1q23.3 sub-band comprises 13 spike genes, with a median spike score of 2.92 (range, 2.24 to 3.83), spanning a 3.94 Mb region from *SLAMF1* to *PBX1*. The 1p31.1 sub-band comprises ten spike genes, with a median spike score of 2.88 (range, 2.14 to 5.68), spanning 13.66 Mb region from *SFRS11* to *TTLL7*, including *IFI44L* and *IFI44*. The 1p36.12 sub-band also comprises ten spike genes, with a median spike score of 2.98 (range, 2.36 to 3.62), spanning 3.45 Mb region from *PLA2G2D* to *ID3*. These genes included *WNT4* and *E2F2*.

Prognostic values of spike (sub)bands using training and test patient cohorts

To look for a prognostic value of a given spike (sub)band, its expression was defined for each patient as the mean of the signals of the spike genes within this (sub)band. Using this parameter, 55 of the 149 spike (sub)bands had prognostic value as determined using the R package MaxStat function and multiple testing correction in the HM cohort. *Online Supplementary Table S8* displays these 55 (sub)bands with good or bad prognosis ranked according to their *P* value using Cox univariate analysis. The list of the 344 spike genes located in these prognostic (sub)bands is also given as well as the known characteristics of the proteins they encode for (*Online Supplementary Table S9*). *Online Supplementary Figure S5* shows the survival Kaplan Meier curves of the four spike (sub)bands with the worst prognosis and the four with the best prognosis. The prognostic information provided by these 55 prognostic spike (sub)bands was summed within a SBS, as defined in the *Design and Methods* section. The SBS had prognostic value when used as a continuous variable ($P \leq 10^{-4}$). Patients of the HM cohort were ranked according to increased SBS, and for a given S value, the difference in survival of patients with a SBS ≤ S or > S was computed. A maximum difference in overall survival was obtained with a SBS cut-point of -103.8, splitting patients into a high-risk group (SBS > -103.8) and a low-risk group (SBS ≤ -103.8). The number of spike prognostic bands could be reduced from 55 to 53 without losing the prognostic power of the SBS. A further decrease in the number of spike prognostic bands down to 28 increased the prognostic value of the score (*P* value decreased by 23%), but with a decrease in the number of patients with a poor prognosis (from 24% to 17%), which is not desirable. Thus, 53 spike prognostic bands were used to build the SBS and are outlined with an asterisk in *Online*

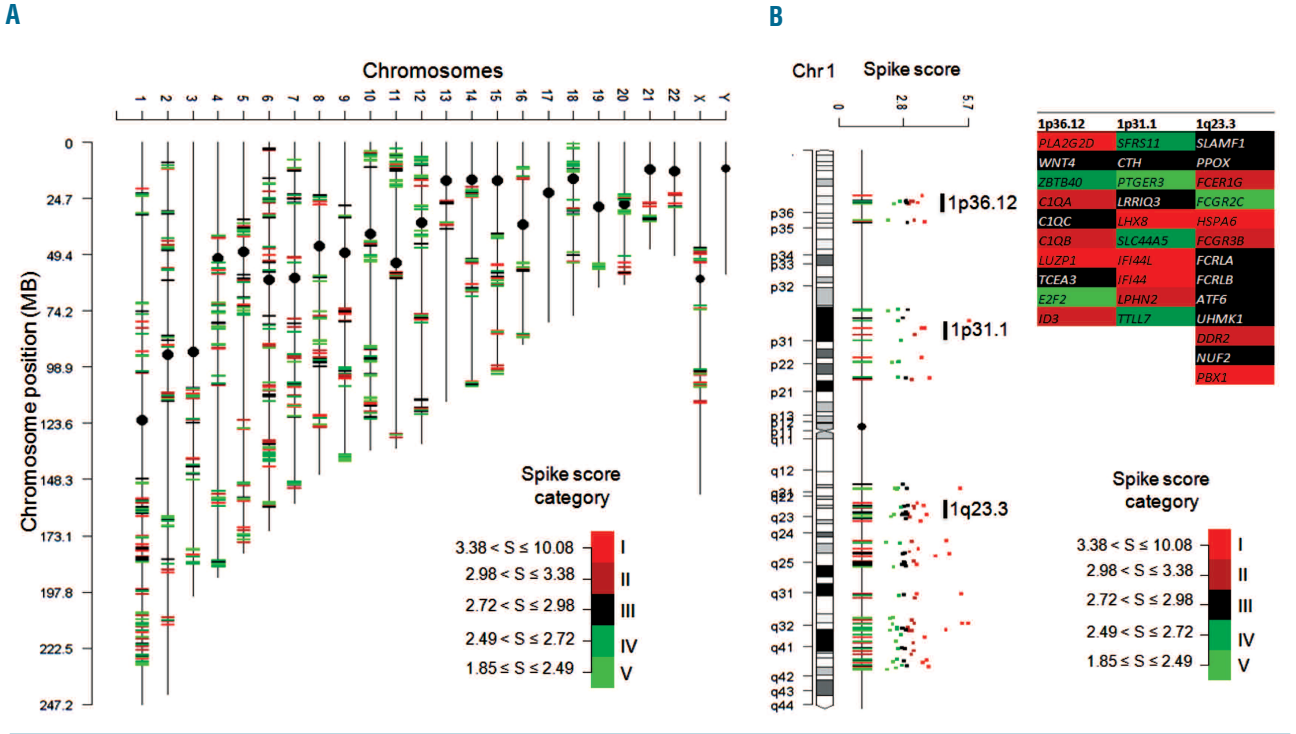


Figure 4. Chromosome distribution of spike genes. (A) Distribution of spike genes in the significant spike (sub)bands. The spike genes included in the 149 spike (sub)bands are displayed along the 24 chromosomes using a color code indicating the spike score category. The black ellipse shows the chromosome centromere. Each bar represents a spike gene. (B) Visualization of the spike (sub)bands and associated spike genes for chromosome 1. The three spike (sub)bands with ≥ ten spike genes are shown. Each bar or dot represents a spike gene.

Supplementary Table S8. A maximum difference in overall survival was obtained with a SBS cut-point of -104.3, splitting patients into a high-risk group of 23.8% patients (SBS > -104.3) with a median overall survival of 27.4 months and a low-risk group of 76.2% patients (SBS ≤ -104.3) in whom the median survival was not reached (Figure 5). The prognostic value of the SBS was compared with standard prognostic variables, including t(4;14), del17p, two previously-published gene expression-based risk scores (UAMS-HRS[17] and IFM[18]) and the gene expression based proliferation index (GPI).[43] By univariate Cox analysis, the SBS, UAMS-HRS, IFM-score and GPI had prognostic value as well as spike MMSET, del17p, β2m, albumin and ISS using the HM patient cohort (*Online Supplementary Table S10*). When compared two by two, only the SBS and del17p remained significant. When these parameters were tested together, only SBS and del17p were prognostic. The strong prognostic value of the SBS in the HM cohort is hardly surprising since the SBS was built using this cohort. It was, therefore, interesting to look at the prognostic value of the score in an independent cohort of 345 patients from UAMS treated with TT2 therapy (UAMS-TT2 cohort). The SBS for each patient of the UAMS-TT2 cohort was computed using parameters defined with the HM cohort, and patients were split into two groups using the percentages defined with the HM cohort. The SBS was prognostic in the UAMS-TT2 cohort. The median overall survival of patients within the high score group was 56 months and not reached in the second group (P<0.0001) (Figure 5). Using Cox univariate analysis, the UAMS-HRS, IFM and GPI scores as well as spike *MMSET* and del17p also had prognostic value. β2m, albumin and ISS could not be evaluated since these data were not publicly available. Comparing these prognostic factors two by two, the SBS remained significant compared to the UAMS-HRS, IFM, GPI, spike *MMSET*, and del17p in the UAMS-TT2 cohort (*Online Supplementary Table S10*). When these parameters were tested together, only HR, spike *MMSET* and del17p were prognostic in the UAMS-TT2 cohort.

Clinical and biological characteristics of patients with high and low spike (sub)band score

Patients age ≥ 65 years, with high β2-microglobulin, low albumin, high C-reactive protein, low hemoglobin, or ISS stages II or III were over-represented among patients with a high-risk SBS (P≤0.05). Other clinical data - Salmon Durie staging, light or heavy chain isotype, occurrence of bone lesions, and serum levels of lactate dehydrogenase - were not significantly different between the SBS high-risk and low-risk groups (*Online Supplementary Table S11*).

Discussion

We have demonstrated that spike genes - *i.e.* genes over-expressed in MM cells from a fraction of newly-diagnosed patients - are significantly clustered in 149 out of the 862 documented bands and sub-bands, which we termed spike (sub)bands. More than one third (55) of these spike (sub)bands are associated with bad (24.2%) or good (12.8%) prognosis using stringent multiple testing correction.

To sum up the prognostic value of these spike (sub)bands, a SBS was defined for each patient, which was prognostic both as a continuous and as a dichotomous variables in two independent cohorts of patients. In multivariate testing, SBS remained independent of del17p in the HM cohort. In this analysis, the two previously reported GEP-based risk scores (UAMS-HRS[17] and IFM[18]), the gene expression based proliferation index,[43] t(4;14), ISS and β2-microglobulin were not independent of del17 and SBS. Comparing these prognostic factors two by two in the UAMS-TT2 cohort, SBS remained significant compared to UAMS-HRS, IFM, GPI,

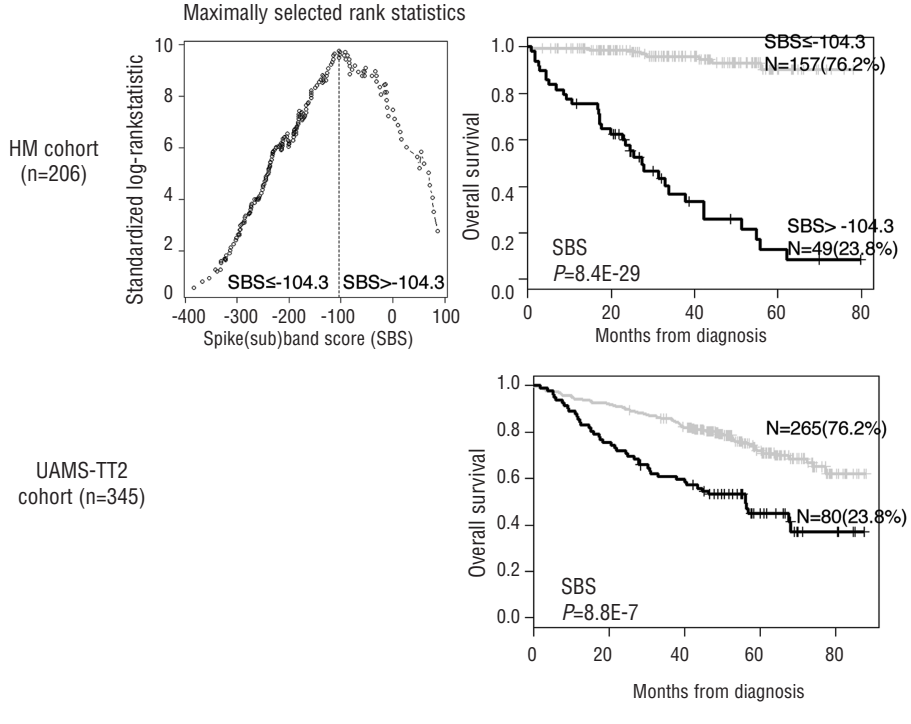


Figure 5. Prognostic value of spike (sub)bands. The prognostic information provided by the 53 prognostic spike (sub)bands was summed within a spike (sub)band score (SBS) as defined in the *Design and Methods* section. Patients of the HM cohort were ranked according to increased SBS and a maximum difference in overall survival was obtained with an SBS of -104.3 splitting patients into high-risk (23.8%) and low-risk (76.2%) groups. The prognostic value of the SBS was tested on an independent cohort of 345 patients from UAMS treated with TT2 therapy (UAMS-TT2 cohort). The parameters to compute the SBS score of patients of the UAMS-TT2 cohort and the proportions delineating the two prognostic groups were those defined with the HM cohort.

spike *MMSET*, and del17p (*Online Supplementary Table S10*). Multivariate analysis of the six prognostic factors together in the UAMS-TT2 cohort showed that UAMS-HRS was independent of del17p and spike *MMSET*. Taking both these results together, these data indicate that there is an interrelation between SBS and UAMS-HRS, *i.e.* part of the prognostic power of the UAMS-HRS is potentially explainable by the expression of spike genes. Thus, the strong prognostic value of the SBS in an independent cohort of patients suggests that gene spiking mechanisms are important pathophysiological mechanisms.

The spiking mechanisms could involve DNA alterations (change in DNA copy number, translocation, mutation), epigenetic regulation, or physiological regulation. Genomic rearrangement is clearly a spiking mechanism, since recurrent MM translocations (involving *FGFR3*, *MMSET*, *CCND1*, *CCND3*, or *MAF* genes and IG loci) are all associated with spike gene expression (in category I, Figure 1). Another line of evidence is that spectral karyotyping in 12 HMCL, identifying complex DNA rearrangements, revealed unique association of spike genes and translocation-associated breakpoints involving the corresponding gene locus. This was illustrated for *CFH*, *EMP1*, and *NID2* genes, whose expression is spiked only in the HMCL with a breakpoint involving the cytoband containing the gene (Figure 2). Of note, no novel translocations involving *IGH* or *IGL* loci in association with spike expression of the gene partners were identified using spectral karyotyping in these 12 HMCL. This suggests that novel translocations involving *IGH* or *IGL* loci are rare. As whole genome sequencing becomes less expensive, this can be further investigated sequencing a large number of primary MM cell samples and HMCL together with spike gene identification. It can also be investigated by selecting probes specific for the three immunoglobulin loci and probes for some of the recurrent spiked genes.

Increase in DNA copy number is not a spiking mechanism. This is not surprising since the algorithm we used to pick up spike genes ([signal ≥ Q3 + 3(Q3-Q1)], Q3 and Q1 being quartiles 3 and 1 of probe set signals among patients' samples) is stringent. A spike signal would not occur with the 1.5-3 fold increase in gene expression expected with an increased DNA copy number. For the same reason there was no link between gender and spike expression of sex chromosome genes. An aberrant expression can be due to a functional upregulation. One example is the proliferation of MM cells, which appears concomitantly to expression of a large number of cell-cycle genes.[44] The spike expression of *CCNE2*, encoding for cyclin E, is illustrative (see *Online Supplementary Figure S3*). Another possibility is genes that are targets of an upstream amplified gene product, a transcription factor for example. This mechanism could be analyzed further using correlation network modeling as exemplified for *MAF* and *MMSET* in Figure 3.

Just over one third of the spike (sub)bands (55 of 149; 36.9%) had prognostic value, despite the use of a stringent Benjamini Hochberg multiple testing correction. Two prognostic groups of MM patients could be predicted with our training HM cohort and were strongly validated with the test UMAS-TT2 cohort. An explanation is that some of the 344 spike genes associated with these prognostic spike (sub)bands encode for proteins modulating tumor survival, proliferation, growth and/or drug resistance. The list of these 344 genes is provided in *Online Supplementary Table S9*. They are equally distributed in the five spike categories displayed in Figure 1 (21.5% in category I, 18.6% in category II, 23.3% in category III, 19.8% in category IV and 16.8% in category V). These genes encode for proteins in various pathways, mainly cell growth and proliferation, cell death, cell development, cell function, maintenance and movement. Some of these 344 spike genes have already been documented as prognostic genes, including cancer testis antigens (*MAGED4*, *XAGE1A*, *SSX2*, *XAGE1A*, *GAGE1*, *GAGE3*) and *CKS1B*.

The method used in this study was developed with Affymetrix microarrays. It is of note that not all Affymetrix probe sets work "properly", *i.e.* show a 100% sensitivity and specificity. This is because Affymetrix probe sets were designed using bioinformatics tools with understandable possible errors in "biological efficacy". In addition, these probe sets cover different parts of the mRNA and work more efficiently when targeting the 3' part because of more efficient reverse transcription. This known imperfection of Affymetrix probe sets is not a major hurdle, providing that biological or clinical markers are used to select and validate them. This is the case for the probe sets used to build the SBS since this score was strongly prognostic in two independent cohorts of patients. There should be no difficulty in applying the current spike method to data obtained with other provider arrays although we could not check this since such data are not publicly available for patients with MM. The current method could be extended to other cancers, in particular to compare the frequency and the location of spike (sub)bands in various cancers, and to look for any associations with cancer type, localization and metastasis.

Authorship and Disclosures

The information provided by the authors about contributions from persons listed as authors and in acknowledgments is available with the full text of this paper at www.haematologica.org.

Financial and other disclosures provided by the authors using the ICMJE (www.icmje.org) Uniform Format for Disclosure of Competing Interests are also available at www.haematologica.org.

References

1. Fonseca R, Bergsagel PL, Drach J, Shaughnessy J, Gutierrez N, Stewart AK, et al. International Myeloma Working Group molecular classification of multiple myeloma: spotlight review. Leukemia. 2009;23 (12):2210-21.
2. Walker BA, Leone PE, Chiecchio L, Dickens NJ, Jenner MW, Boyd KD, et al. A compendium of myeloma-associated chromosomal copy number abnormalities and their prognostic value. Blood. 2010;116(15): e56-65.
3. Avet-Loiseau H, Malard F, Campion L, Magrangeas F, Sebban C, Lioure B, et al. Translocation t(14;16) and multiple myeloma: is it really an independent prognostic factor? Blood. 2011;117(6):2009-11.
4. Hanamura I, Stewart JP, Huang Y, Zhan F, Santra M, Sawyer JR, et al. Frequent gain of chromosome band 1q21 in plasma-cell dyscrasias detected by fluorescence in situ hybridization: incidence increases from MGUS to relapsed myeloma and is related to prognosis and disease progression fol-

lowing tandem stem-cell transplantation. Blood. 2006;108(5):1724-32.
5. Liu P, Leong T, Quam L, Billadeau D, Kay NE, Greipp P, et al. Activating mutations of N- and K-ras in multiple myeloma show different clinical associations: analysis of the Eastern Cooperative Oncology Group phase III trial. Blood. 1996;88(7):2699-706.
6. Chng WJ, Gonzalez-Paz N, Price-Troska T, Jacobus S, Rajkumar SV, Oken MM, et al. Clinical and biological significance of RAS mutations in multiple myeloma. Leukemia. 2008;22(12):2280-4.
7. Keats JJ, Fonseca R, Chesi M, Schop R, Baker A, Chng WJ, et al. Promiscuous mutations activate the noncanonical NF-kappaB pathway in multiple myeloma. Cancer Cell. 2007;12(2):131-44.
8. Annunziata CM, Davis RE, Demchenko Y, Bellamy W, Gabrea A, Zhan F, et al. Frequent engagement of the classical and alternative NF-kappaB pathways by diverse genetic abnormalities in multiple myeloma. Cancer Cell. 2007;12(2):115-30.
9 . Carrasco DR, Tonon G, Huang Y, Zhang Y, Sinha R, Feng B, et al. High-resolution genomic profiles define distinct clinicopathogenetic subgroups of multiple myeloma patients. Cancer Cell. 2006;9(4):313-25.
10. Avet-Loiseau H, Li C, Magrangeas F, Gouraud W, Charbonnel C, Harousseau JL, et al. Prognostic significance of copy-number alterations in multiple myeloma. J Clin Oncol. 2009;27(27):4585-90.
11. Crawley JJ, Furge KA. Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. Genome Biol. 2002;3(12): RESEARCH0075.
12. Santarius T, Shipley J, Brewer D, Stratton MR, Cooper CS. A census of amplified and overexpressed human cancer genes. Nat Rev Cancer. 2010;10(1):59-64.
13. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177-83.
14. Broyl A, Hose D, Lokhorst H, de Knegt Y, Peeters J, Jauch A, et al. Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. Blood. 2010;116(14):2543-53.
15. Zhan F, Huang Y, Colla S, Stewart JP, Hanamura I, Gupta S, et al. The molecular classification of multiple myeloma. Blood. 2006;108(6):2020-8.
16. Bergsagel PL, Kuehl WM. Molecular pathogenesis and a consequent classification of multiple myeloma. J Clin Oncol. 2005;23 (26):6333-8.
17. Shaughnessy JD, Jr., Zhan F, Burington BE, Huang Y, Colla S, Hanamura I, et al. A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. Blood. 2007;109(6):2276-84.
18. Decaux O, Lode L, Magrangeas F, Charbonnel C, Gouraud W, Jezequel P, et al. Prediction of survival in multiple myeloma based on gene expression profiles reveals cell cycle and chromosomal instability signatures in high-risk patients and hyperdiploid signatures in low-risk patients: a study of the Intergroupe Francophone du Myelome. J Clin Oncol. 2008;26 (29):4798-805.
19. Hose D, Reme T, Meissner T, Moreaux J, Seckinger A, Lewis J, et al. Inhibition of aurora kinases for tailored risk-adapted treatment of multiple myeloma. Blood. 2009;113(18):4331-40.
20. Seckinger A, Meissner T, Moreaux J, Goldschmidt H, Fuhler GM, Benner A, et al. Bone morphogenic protein 6: a member of a novel class of prognostic factors expressed by normal and malignant plasma cells inhibiting proliferation and angiogenesis. Oncogene. 2009;28(44):3866-79.
21. Sprynski AC, Hose D, Caillot L, Reme T, Shaughnessy JD, Jr., Barlogie B, et al. The role of IGF-1 as a major growth factor for myeloma cell lines and the prognostic relevance of the expression of its receptor. Blood. 2009;113(19):4614-26.
22. Chesi M, Bergsagel PL, Brents LA, Smith CM, Gerhard DS, Kuehl WM. Dysregulation of cyclin D1 by translocation into an IgH gamma switch region in two multiple myeloma cell lines [see comments]. Blood. 1996;88:674-81.
23. Chesi M, Nardini E, Lim RS, Smith KD, Kuehl WM, Bergsagel PL. The t(4;14) translocation in myeloma dysregulates both FGFR3 and a novel gene, MMSET, resulting in IgH/MMSET hybrid transcripts. Blood. 1998;92(9):3025-34.
24. Fonseca R, Barlogie B, Bataille R, Bastard C, Bergsagel PL, Chesi M, et al. Genetics and cytogenetics of multiple myeloma: a workshop report. Cancer Res. 2004;64(4):1546-58.
25. Goldschmidt H, Sonneveld P, Cremer FW, van der Holt B, Westveer P, Breitkreutz I, et al. Joint HOVON-50/GMMG-HD3 randomized trial on the effect of thalidomide as part of a high-dose therapy regimen and as maintenance treatment for newly diagnosed myeloma patients. Ann Hematol. 2003;82(10):654-9.
26. Moreaux J, Klein B, Bataille R, Descamps G, Maiga S, Hose D, et al. A high-risk signature for patients with multiple myeloma established from the molecular classification of human myeloma cell lines. Haematologica. 2011;96(4):574-82.
27. Barlogie B, Tricot G, Rasmussen E, Anaissie E, van Rhee F, Zangari M, et al. Total therapy 2 without thalidomide in comparison with total therapy 1: role of intensified induction and posttransplantation consolidation therapies. Blood. 2006;107(7):2633-8.
28. Xiong W, Wu X, Starnes S, Johnson SK, Haessler J, Wang S, et al. An analysis of the clinical and biologic significance of TP53 loss and the identification of potential novel transcriptional targets of TP53 in multiple myeloma. Blood. 2008;112(10): 4235-46.
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.
30. Neben K, Jauch A, Bertsch U, Heiss C, Hielscher T, Seckinger A, et al. Combining information regarding chromosomal aberrations t(4;14) and del(17p13) with the International Staging System classification allows stratification of myeloma patients undergoing autologous stem cell transplantation. Haematologica. 2010;95(7):1150-7.
31. Popp S, Jauch A, Schindler D, Speicher MR, Lengauer C, Donis-Keller H, et al. A strategy for the characterization of minute chromosome rearrangements using multiple color fluorescence in situ hybridization with chromosome-specific DNA libraries and YAC clones. Hum Genet. 1993;92 (6):527-32.
32. Cremer FW, Bila J, Buck I, Kartal M, Hose D, Ittrich C, et al. Delineation of distinct subgroups of multiple myeloma and a model for clonal evolution based on interphase cytogenetics. Genes Chromosomes Cancer. 2005;44(2):194-203.
33. Hose D, Rossi JF, Ittrich C, DeVos J, Benner A, Reme T, et al. A New molecular classification of multiple myeloma using gene expression profiling and fluorescence in situ hybridisation as predictor for event free survival. Blood. 2004;104(11):Abstract 73.
34. Wuilleme S, Robillard N, Lode L, Magrangeas F, Beris H, Harousseau JL, et al. Ploidy, as detected by fluorescence in situ hybridization, defines different subgroups in multiple myeloma. Leukemia. 2005;19 (2):275-8.
35. Rao PH, Cigudosa JC, Ning Y, Calasanz MJ, Iida S, Tagawa S, et al. Multicolor spectral karyotyping identifies new recurring breakpoints and translocations in multiple myeloma. Blood. 1998;92(5):1743-8.
36. Schrock E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, et al. Multicolor spectral karyotyping of human chromosomes. Science. 1996;273(5274): 494-7.
37. Sawyer JR, Lukacs JL, Munshi N, Desikan KR, Singhal S, Mehta J, et al. Identification of new nonrandom translocations in multiple myeloma with multicolor spectral karyotyping. Blood. 1998;92(11):4269-78.
38. Telenius H, Pelmear AH, Tunnacliffe A, Carter NP, Behmel A, Ferguson-Smith MA, et al. Cytogenetic analysis by chromosome painting using DOP-PCR amplified flow-sorted chromosomes. Genes Chromosomes Cancer. 1992;4(3):257-63.
39. Shaffer LG TN, eds. ISCN 2005. An International System for Human Cytogenetic Nomenclature. Basel, Switzerland: S Karger. 2005.
40. Reme T, Hose D, De Vos J, Vassal A, Poulain PO, Pantesco V, et al. A new method for class prediction based on signed-rank algorithms applied to Affymetrix microarray experiments. BMC Bioinformatics. 2008;9: 16.
41. Assou S, Le Carrour T, Tondeur S, Strom S, Gabelle A, Marty S, et al. A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. Stem Cells. 2007;25(4):961-73.
42. Team RDC. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.
43. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
44. Hose D, DeVos J, Muller N, Rossi JF, Heib C, Mahtouk K, et al. Clonal and subclonal cytogenetic aberrations: association with D-type cyclin expression and event-free survival (EFS) in multiple myeloma (MM). Blood. 2006;108(11):982a-a.
45. Hose D, Reme T, Hielscher T, Moreaux J, Meissner T, Seckinger A, et al. Proliferation is a central independent prognostic factor and target for personalized and risk adapted treatment in multiple myeloma. Haematologica. 2011;96(1):87-95.
46. Hose D, Reme T, Hielscher T, Moreaux J, Meissner T, Seckinger A, et al. A gene expression based proliferation index as independent prognostic factor in multiple myeloma. Blood. 2008;112(11):589.