The Reconstruction of Transcriptional Networks Reveals Critical Genes with Implications for Clinical Outcome of Multiple Myeloma

Luca Agnelli¹, Mattia Forcato², Francesco Ferrari², Giacomo Tuana¹, Katia Todoerti¹, Brian A. Walker³, Gareth J. Morgan³, Luigia Lombardi¹, Silvio Bicciato², and Antonino Neri¹

Abstract

Purpose: The combined use of microarray technologies and bioinformatics analysis has improved our understanding of the biological complexity of multiple myeloma (MM). In contrast, the application of the same technology in the attempt to predict clinical outcome has been less successful with the identification of heterogeneous molecular signatures. Herein, we have reconstructed gene regulatory networks in a panel of 1,883 samples from MM patients derived from publicly available gene expression sets, to allow the identification of robust and reproducible signatures associated with poor prognosis across independent data sets.

Experimental Design: Gene regulatory networks were reconstructed by using Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe) and microarray data from seven MM data sets. Critical analysis of network components was applied to identify genes playing an essential role in transcriptional networks, which are conserved between data sets.

Results: Network critical analysis revealed that (i) CCND1 and CCND2 were the most critical genes; (ii) CCND2, AIF1, and BLNK had the largest number of connections shared among the data sets; (iii) robust gene signatures with prognostic power were derived from the most critical transcripts and from shared primary neighbors of the most connected nodes. Specifically, a critical-gene model, comprising FAM53B, KIF21B, WHSC1, and TMPO, and a neighbor-gene model, comprising BLNK shared neighbors CSGALNACT1 and SLC7A7, predicted survival in all data sets with follow-up information.

Conclusions: The reconstruction of gene regulatory networks in a large panel of MM tumors defined robust and reproducible signatures with prognostic importance, and may lead to identify novel molecular mechanisms central to MM biology. Clin Cancer Res; 17(23); 1–11. ©2011 AACR.

Introduction

In the past decade, a large amount of whole-genome transcriptional profiling data have been generated for almost all tumor types, among which multiple myeloma (MM) is characterized by one of the largest publicly available data sets. To date, several studies aimed at stratifying MM samples have been published, based on either molecular features (1–3) or transcriptional profiles (4–6). However, different microarray studies have identified different signatures able to predict clinical outcome. Unfortunately, the minimal number of overlapping genes between the signatures raised concerns about both their biological and clinical relevance, and we are far from a consensus in the definition of transcriptional prognostic markers for high-risk MM. The skepticism about reliability and reproducibility of gene signatures for prognostic marker discovery stems from the complexity of microarray data, the limited number of samples analyzed, and the computational approach used. These limitations can be overcome by the integration of multiple, independently generated data sets and their analysis, using methods that allow the reverse engineering and reconstruction of regulatory networks.

Recently, the Algorithm for Reconstruction of Accurate Cellular Network (ARACNe) was successfully applied in a large data set of B-cell tumors expression profiles and identified c-MYC and BCL6 as critical genes (7–9). This type of analysis is useful for the identification of previously...
unknown targets (namely, neighbor genes, i.e., genes directly connected to another gene) that can be subsequently validated experimentally. Such direct regulatory interactions are likely identified as those mediated by a transcription factor (TF) binding to a target gene’s promoter region, although other types of interactions (e.g., for transcript coding proteins involved in stable complex formation) may also be identified. Moreover, network theory-derived procedures have been developed that describe the internal stability of complex networks as resilience to perturbation, which can be used to identify genes that have a critical role in the stability and functioning of the network. This approach allows to define master regulator genes that might explain the biology of disease subsets (11–13). The selection of such genes, based on their role in a critical regulatory network, rather than solely on their level of expression, may improve the reliability and robustness of expression analysis to predict outcome. This approach has been shown as useful in other tumors, including hematologic malignancies as chronic lymphocytic leukemia (14).

In this study, we (i) reconstructed gene regulatory networks in MM by applying ARACNe to the gene expression profiles of 1,883 samples included in 7 publicly available data sets; (ii) investigated the structure of the networks to identify critical genes and infer novel regulatory relationships; and (iii) defined robust gene signatures with prognostic power from the most critical genes and the commonly shared neighbors.

Materials and Methods

The entire workflow for network reconstruction, identification of critical genes and design of robust prognostic signatures is represented in Supplementary Fig. S1. R software was used for all calculations with appropriate functions.

Network reconstruction using ARACNe

The gene expression data were obtained from a proprietary (GSE13591 at NCBI GEO repository, http://www.ncbi.nlm.nih.gov/geo/; ref. 15) and 6 publicly available large data sets (GSE755, GSE2658, GSE6477, GSE9782, GSE15695, GSE19784) including plasma cells from healthy donors, monoclonal gammopathy of undetermined significance, MM, and plasma cell leukemia (3, 6, 16–19). Three data sets (GSE2658, GSE9782, and GSE15695) included outcome information in terms of overall survival (OS). With the exception of GSE2658 and GSE9782 series, expression values were extracted from cell intensity files, using GeneAnnot custom chip definition files, and normalized by Robust Multi-array Average procedure, as previously described (15). In the case of GSE2658 and GSE9782 series, only the expression values calculated by using Affymetrix MAS5.0 algorithm were available. Because the critical analysis of network components was developed for gene-centered data, in those 2 data sets we assigned a unique value (quantified as the median signal over multiple probe sets; probe sets that have a Pearson correlation with the median profile lower than 0.5 were discarded from the calculation of the median signal) to all Entrez Gene IDs represented by multiple probe sets. Conversely, in the following survival analysis, to avoid biases in the comparison with previously published results, the gene expression signals of GSE15695 were quantified by using Affymetrix MAS5.0 algorithm and GSE2658 and GSE9782 data were used as downloaded. We reconstructed gene regulatory networks by using ARACNe, as summarized in the protocol by Margolin and colleagues (Fig. 4 in ref. 20). Briefly, the algorithm first uses the expression data to calculate pairwise mutual information (MI), a dimensionless quantity that represents the mutual dependence between 2 transcripts, through a computationally efficient Gaussian kernel estimator. The parameters of the kernel width and MI threshold were calculated by R scripts (generated from the original MATLAB procedures), under standard conditions. The P value to determine the MI threshold and the data processing inequality (DPI) tolerance were set to $1 \times 10^{-7}$ and 10%, respectively. Through the DPI tolerance, any regulatory loop formed by 3 interacting genes would result in the weakest of the 3 interactions being removed, unless its strength is within 10% of that of the intermediate interaction (20). A list of probe sets for the TFs represented in HG-U95Av2, HG-U133A, and HG-U133 Plus 2.0 platforms was used to prevent the DPI from removing transcriptional interactions in favor of nontranscriptional ones (interactions between 2 non-TFs). In any data set, genes with low signal variability across samples were eliminated by using an entropy-based filter (21), leading to 20% of the transcripts being discarded. The network of inferred regulatory interactions among genes was interactively browsed using Cytoscape (22) and the expression levels of the selected genes were visualized using the SpotXplore plug-in.
(23). The fully reconstructed networks were supplied as Supplementary Data in the form of Cytoscape format-
files, available at our Institution web site (http://
www.xlab.unimo.it/MM_Network/).

Critical analysis of network components and global criticality score calculation

The procedure for critical analysis calculation was
implemented in C++ and carried out by measuring the
drop in the network performance caused by single-node
removal. As previously described (10, 24), network per-
formance is defined as the network efficiency, a measure
that quantifies how efficiently the network nodes
exchange information. Specifically, after the quantifica-
tion of network efficiency, for each gene/node i the
network efficiency was requantified after removal of the
node and of its adjacent edges, and the difference in
network efficiency before and after removal of the node
i is defined as criticality. Each gene has been ranked in
terms of criticality and a global node criticality score (GCS)
has been defined to quantify the level of criticality of each
gene across all data sets, for all those nodes that are
common to all networks. In the quantification of GCS,
genes are first ranked according to criticality and degree
(namely, the number of interactions for each gene) in
each data set. Comparing these ranks is equivalent to
calculating the degree ranking (i.e., in a noncritical net-
work), the rank of a gene/node i is identified by node degree
(ki) and degree value partly independent from its degree and, in partic-
ular, if ki > max, then the node is critical without being
highly connected, conversely the node is less critical than
simply expected from the number of its connections.
Degrees k’ and ki are used to calculate the difference
Δki = (k’ - ki)/kmmax that quantifies, for a given criticality
rank ri, how many degrees a node (gene) has gained or
lost in the critical network as compared with the noncrit-
ical one. The difference was scaled by the maximum
degree kmmax of any network to allow a fair comparison
among genes of networks generated from data obtained by
different types of microarrays. The calculations of
GCS was repeated so that each of the 7 data sets was excluded
at a time: the final GCS represents the median Δki of
7 partial GCS (thus calculated from 6 of the 7 data sets,
alternatively), for each gene.

Survival analysis

For survival analysis, Cox models have been designed
by using each considered gene lists and patients included
in the TT2 trial of GSE2658 as training set (3). Data from
GSE9782 (19) and GSE15695 (6) were used as indepen-
dent test set. Gene lists were derived from the top 100
critical nodes and from primary neighbors, shared among
multiple data sets, of the most connected nodes. For each
node, the corresponding list of Affymetrix probe sets was
retrieved. A Cox model was designed with each list to
search for a correlation between the gene expression levels
and survival on training set samples. In the case of a
significant association with outcome, each list was
reduced to derive a smaller prognostic gene signature,
using a Cox proportional hazards model in the globaltest
function of R software. Briefly, the globaltest is based on
regression models in which the distribution of the
response variable is modeled as a function of the covar-
iates (25). In this context, the probe set expression values
correspond to the covariates, the regression model is the
Cox proportional hazards model, and the response
variable is survival. The decomposition of the test statistic
into the contributions of each covariate, allows identi-
ﬁcation of those covariates (or groups of covariates) that
correlate most to a significant test result. All of the combi-
inations of the probe sets, or fraction of those, that
reached signiﬁcant levels where tested for correlation with
OS, and the signiﬁcance of each combination was adjust-
ed for multiple testing, using Benjamini and Hochberg
procedure. To test the correlation of each combination
with OS, patients were then stratified into 2 groups by
using the gene signatures and hierarchical clustering with
Euclidean and Ward as distance and linkage metrics,
respectively; where only 1 probe set was tested, the
median of the expression values was chosen to generate
the 2 groups. The groups identified by this approach were
tested for association with survival by using the Kaplan–
Meier estimator and log-rank test, and P values were
calculated according to the standard normal asymptotic
distribution. Survival analysis was conducted with the
km.coxph.plot function of the survcomp R package.
Independence between commonly used prognostic factors
and gene signatures was tested by using multivariate Cox
regression procedure of survival R package. Speciﬁcally, to
assess whether the gene signatures add additional prog-
nostic information over clinical variables, therapy, or
outcome classiﬁcations, Cox proportional hazards mod-
els were constructed by using any single predictor alone
and with the gene signature stratification, and testing the
difference between the 2 models against a χ² distribution
with 1 degree of freedom (ANOVA).

The performance of the critical and neighbor-gene sig-
natures has been compared with the Intergroupe Franc-
phone du Myelome (IFM) 15-gene (4), the University of
Arkansas for Medical Sciences (UAMS) 17-gene (5), and
the Myeloma IX 6-gene signatures (6). In particular, the IFM 15-
gene model was applied ranking the patients according
to the risk score described in Decaux and colleagues (4) and
classifying them into risk groups using the quartiles. Cus-
tom probe sets deﬁning the score have been converted into
Affymetrix HG-U133 Plus 2.0 probe sets using their Table
A12 of the Supplementary Information. In the UAMS 17-
gene model, we classified each patient using discriminant
score and cutoff values described by Shaughnessy and colleagues (3), whereas for the 6-gene model, we used the same classification presented by Dickens and colleagues (6).

**Results**

**Reconstruction of plasma cell dyscrasia transcriptional networks**

ARACNe was used to reconstruct cellular networks from a set of global gene expression profiles including normal and myelomatous plasma cells derived from 7 publicly available series, comprising a total of 1,883 cases. Table 1 and Supplementary Fig. S2 summarize the metrics and the topologic characteristics of the transcriptional networks reconstructed by using data from the various series. Overall, networks derived from the same type of microarray shared similar behaviors, particularly with regard to the total amount of nodes (approximately 10,000 for HG-U133A and more than 15,000 for HG-U133 Plus 2.0): the networks had comparable probability distribution of connectivity degrees and nodes tended to connect to nodes with a similar connectivity.

**Identification of critical nodes**

Analysis of network critical components was then used to assess the importance of genes as putative regulatory elements. The critical nodesgenes of each network were determined by removing individual genes from the network and calculating the network efficiency after removal. We found that genes with relatively low numbers of connections impaired network stability in all the data sets analyzed. This finding is consistent with previous evidence suggesting that reverse engineering approaches are able to discover master regulator genes with limited numbers of interactions largely affecting the transcriptome (11, 12, 26). Table 2 and Supplementary Table S1 list the top 100 genes with the highest GCS derived from our analysis (the whole criticality scores for each data set are provided as Supplementary Table S2). Notably CCND1 and CCND2 emerged as the most critical genes across all data sets and WHSC1 and FGER3 genes deregulated by (4;14) translocation were also identified.

A critical-gene model independently predicts overall survival

On the basis of their shared relevance within transcriptional networks, the 100 genes identified have characteristics that make them good candidates as being central to the clinical outcome of MM. Thus, we went on to test whether their expression was associated with OS in the 3 data sets with associated prognostic information (4–6), including 351, 264, and 247 samples, respectively. In each cohort, samples were first classified into 2 groups on the basis of the expression levels of the 219 HG-U133A probe sets (Supplementary Table S3) matching the top 100 critical genes; then, the 2 groups were compared by using the log-rank test to evaluate the prognostic value of the gene model. This analysis revealed a significant correlation between expression levels and OS in 2 of the 3 independent data sets (Fig. 1).

To further investigate the predictive power of the signature genes, the significance of the correlation between the expression level of each probe set and survival was tested. Specifically, a proportional hazards model for any probe set was built by using the TT2 samples of GSE2658 as training set, quantifying the sign (positive or negative) and the significance of any correlation found with survival (Supplementary Table S3), and then testing the correlation with OS in the 2 test sets GSE9782 and GSE15695. This approach identified 25 significant probe sets, matching 19 unique genes, at $P < 0.05$ (Fig. 2A). However, when applied to test sets, none of these probe sets alone was able to predict OS in both the test cohorts and consequently we went on to define a reduced gene signature on the basis of all the probe sets with $P < 0.01$ in the proportional hazards models. These probe sets were FAM53B (203206_at), positively correlated with survival, KIF21B (204411_at), WHSC1 (209052_s_at and 209053_s_at), TMPO (209753_s_at), and TXNDC13 (201580_s_at) negatively correlated. The expression levels of the genes in this reduced signature (comprising 5 probe sets, after removing the less significant WHSC1 209053_s_at to avoid redundancy) were used to cluster test set samples into 2 groups, using all of the possible combinations of the 5 probe sets. These 2 groups were associated with statistically different survival probabilities in all 3 data sets.

**Table 1. Metrics of the reconstructed networks**

<table>
<thead>
<tr>
<th>Data sets</th>
<th>No. of samples</th>
<th>Array (CEL available)</th>
<th>Nodes</th>
<th>Edges</th>
<th>$k_{max}$</th>
<th>$k_{mean}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE755</td>
<td>173</td>
<td>HG-U95Av2 (Y)</td>
<td>6,683</td>
<td>57,022</td>
<td>148</td>
<td>17.04</td>
</tr>
<tr>
<td>GSE2658</td>
<td>559</td>
<td>HG-U133 Plus 2.0 (N)</td>
<td>16,195</td>
<td>115,726</td>
<td>1,098</td>
<td>14.3</td>
</tr>
<tr>
<td>GSE6477</td>
<td>162</td>
<td>HG-U133A (Y)</td>
<td>9,620</td>
<td>91,710</td>
<td>301</td>
<td>19.07</td>
</tr>
<tr>
<td>GSE9782</td>
<td>264</td>
<td>HG-U133A (N)</td>
<td>10,327</td>
<td>75,289</td>
<td>1,071</td>
<td>14.58</td>
</tr>
<tr>
<td>GSE13591a</td>
<td>158</td>
<td>HG-U133A (Y)</td>
<td>9,666</td>
<td>86,846</td>
<td>219</td>
<td>17.97</td>
</tr>
<tr>
<td>GSE15695</td>
<td>247</td>
<td>HG-U133 Plus 2.0 (Y)</td>
<td>15,087</td>
<td>130,966</td>
<td>622</td>
<td>17.36</td>
</tr>
<tr>
<td>GSE19784</td>
<td>320</td>
<td>HG-U133 Plus 2.0 (Y)</td>
<td>15,111</td>
<td>125,296</td>
<td>488</td>
<td>16.58</td>
</tr>
</tbody>
</table>

**NOTE:** $k$, degree.

Abbreviation: CEL, cell intensity file.

*Described as MM158GE data set in the abstract communication from Biasiolo and colleagues (24).
### Table 2. Top 100 common critical nodes ranked by GCS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>GCS</th>
<th>#fn</th>
<th>%TN</th>
<th>Symbol</th>
<th>GCS</th>
<th>#fn</th>
<th>%TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCND2</td>
<td>0.0513</td>
<td>14</td>
<td>6.36</td>
<td>GPX3</td>
<td>0.0136</td>
<td>1</td>
<td>1.37</td>
</tr>
<tr>
<td>CCND1</td>
<td>0.0486</td>
<td>4</td>
<td>2.15</td>
<td>CST6</td>
<td>0.0135</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>MS4A1</td>
<td>0.0444</td>
<td>1</td>
<td>0.72</td>
<td>DOCK3</td>
<td>0.0133</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>MND4</td>
<td>0.0413</td>
<td>3</td>
<td>3.53</td>
<td>FBLN2</td>
<td>0.0133</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>AIF1</td>
<td>0.0363</td>
<td>14</td>
<td>5.88</td>
<td>DDR1</td>
<td>0.0132</td>
<td>3</td>
<td>2.38</td>
</tr>
<tr>
<td>CLEC11A</td>
<td>0.0315</td>
<td>1</td>
<td>0.78</td>
<td>NFATC1</td>
<td>0.0131</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>ST14</td>
<td>0.0295</td>
<td>7</td>
<td>6.25</td>
<td>LAPT M5</td>
<td>0.0131</td>
<td>6</td>
<td>4.62</td>
</tr>
<tr>
<td>MAPK8IP3</td>
<td>0.0291</td>
<td>0</td>
<td>0.00</td>
<td>SYK</td>
<td>0.0130</td>
<td>1</td>
<td>1.33</td>
</tr>
<tr>
<td>WSCD1</td>
<td>0.0290</td>
<td>2</td>
<td>1.74</td>
<td>GULP1</td>
<td>0.0130</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>ROR2</td>
<td>0.0272</td>
<td>0</td>
<td>0.00</td>
<td>CD19</td>
<td>0.0129</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>KIF21B</td>
<td>0.0269</td>
<td>5</td>
<td>3.73</td>
<td>LAG3</td>
<td>0.0129</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>WHSC1</td>
<td>0.0266</td>
<td>5</td>
<td>3.27</td>
<td>CCT2</td>
<td>0.0129</td>
<td>6</td>
<td>2.34</td>
</tr>
<tr>
<td>RUNX1</td>
<td>0.0246</td>
<td>0</td>
<td>0.00</td>
<td>PDZK1P1</td>
<td>0.0128</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>LMNA</td>
<td>0.0213</td>
<td>2</td>
<td>1.69</td>
<td>TMPO</td>
<td>0.0126</td>
<td>2</td>
<td>1.25</td>
</tr>
<tr>
<td>PLCG2</td>
<td>0.0209</td>
<td>2</td>
<td>1.55</td>
<td>FMR1</td>
<td>0.0125</td>
<td>3</td>
<td>2.11</td>
</tr>
<tr>
<td>FCER2</td>
<td>0.0208</td>
<td>2</td>
<td>2.15</td>
<td>STEAP1</td>
<td>0.0125</td>
<td>1</td>
<td>0.89</td>
</tr>
<tr>
<td>P2RY6</td>
<td>0.0203</td>
<td>0</td>
<td>0.00</td>
<td>VPR2</td>
<td>0.0123</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>CHST7</td>
<td>0.0194</td>
<td>1</td>
<td>0.95</td>
<td>PALLD</td>
<td>0.0122</td>
<td>1</td>
<td>0.76</td>
</tr>
<tr>
<td>CD33</td>
<td>0.0190</td>
<td>1</td>
<td>1.52</td>
<td>NISMAF</td>
<td>0.0120</td>
<td>1</td>
<td>1.25</td>
</tr>
<tr>
<td>FGFR3</td>
<td>0.0186</td>
<td>2</td>
<td>1.75</td>
<td>SCAMP5</td>
<td>0.0120</td>
<td>2</td>
<td>2.25</td>
</tr>
<tr>
<td>USP1</td>
<td>0.0186</td>
<td>2</td>
<td>0.98</td>
<td>NCA M1</td>
<td>0.0118</td>
<td>3</td>
<td>2.75</td>
</tr>
<tr>
<td>MUC1</td>
<td>0.0174</td>
<td>2</td>
<td>1.52</td>
<td>PRKD2</td>
<td>0.0117</td>
<td>4</td>
<td>2.48</td>
</tr>
<tr>
<td>PRAME</td>
<td>0.0163</td>
<td>0</td>
<td>0.00</td>
<td>ATP6V0E2</td>
<td>0.0117</td>
<td>1</td>
<td>1.23</td>
</tr>
<tr>
<td>TGFA</td>
<td>0.0163</td>
<td>0</td>
<td>0.00</td>
<td>VEGFC</td>
<td>0.0117</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>CDH2</td>
<td>0.0162</td>
<td>1</td>
<td>1.04</td>
<td>TESK1</td>
<td>0.0116</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>TNFSF8</td>
<td>0.0162</td>
<td>0</td>
<td>0.00</td>
<td>OSBPL1A</td>
<td>0.0116</td>
<td>4</td>
<td>4.35</td>
</tr>
<tr>
<td>LAMA2</td>
<td>0.0162</td>
<td>0</td>
<td>0.00</td>
<td>SF X3</td>
<td>0.0115</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>NACC2</td>
<td>0.0160</td>
<td>0</td>
<td>0.00</td>
<td>FZD2</td>
<td>0.0114</td>
<td>1</td>
<td>1.14</td>
</tr>
<tr>
<td>TRIO</td>
<td>0.0159</td>
<td>1</td>
<td>1.08</td>
<td>FANCL</td>
<td>0.0114</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>HEG1</td>
<td>0.0158</td>
<td>1</td>
<td>1.45</td>
<td>PKP4</td>
<td>0.0114</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>NCF4</td>
<td>0.0158</td>
<td>2</td>
<td>2.08</td>
<td>MYBPC2</td>
<td>0.0114</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>PTPR2</td>
<td>0.0158</td>
<td>0</td>
<td>0.00</td>
<td>ZBTB16</td>
<td>0.0114</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>FADS1</td>
<td>0.0156</td>
<td>1</td>
<td>1.25</td>
<td>GCAT</td>
<td>0.0113</td>
<td>1</td>
<td>1.32</td>
</tr>
<tr>
<td>FCGBP</td>
<td>0.0153</td>
<td>0</td>
<td>0.00</td>
<td>COL4A6</td>
<td>0.0112</td>
<td>1</td>
<td>1.10</td>
</tr>
<tr>
<td>MAPKAPK5</td>
<td>0.0152</td>
<td>1</td>
<td>0.96</td>
<td>ZFP36L2</td>
<td>0.0112</td>
<td>2</td>
<td>2.33</td>
</tr>
<tr>
<td>ITGA2B</td>
<td>0.0151</td>
<td>4</td>
<td>2.96</td>
<td>FLNB</td>
<td>0.0109</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>TGFB2</td>
<td>0.0150</td>
<td>1</td>
<td>0.85</td>
<td>TMX4</td>
<td>0.0109</td>
<td>1</td>
<td>0.65</td>
</tr>
<tr>
<td>FAM53B</td>
<td>0.0148</td>
<td>0</td>
<td>0.00</td>
<td>PLXNB2</td>
<td>0.0109</td>
<td>1</td>
<td>1.06</td>
</tr>
<tr>
<td>GDPD5</td>
<td>0.0148</td>
<td>1</td>
<td>0.79</td>
<td>C2</td>
<td>0.0109</td>
<td>2</td>
<td>0.00</td>
</tr>
<tr>
<td>RASGRP1</td>
<td>0.0148</td>
<td>9</td>
<td>7.32</td>
<td>GP1BA</td>
<td>0.0108</td>
<td>1</td>
<td>0.49</td>
</tr>
<tr>
<td>FAM171A1</td>
<td>0.0146</td>
<td>2</td>
<td>1.85</td>
<td>LHFL2</td>
<td>0.0107</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>SV2B</td>
<td>0.0145</td>
<td>0</td>
<td>0.00</td>
<td>SERPINI1</td>
<td>0.0106</td>
<td>6</td>
<td>4.51</td>
</tr>
<tr>
<td>PIK3R3</td>
<td>0.0145</td>
<td>2</td>
<td>1.67</td>
<td>TNFAIP2</td>
<td>0.0106</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>ANK3</td>
<td>0.0143</td>
<td>1</td>
<td>0.93</td>
<td>CLDN3</td>
<td>0.0106</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>BLNK</td>
<td>0.0143</td>
<td>11</td>
<td>6.40</td>
<td>ANAPC13</td>
<td>0.0106</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>CEACAM1</td>
<td>0.0142</td>
<td>1</td>
<td>1.18</td>
<td>RAB3B</td>
<td>0.0106</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>SPN</td>
<td>0.0139</td>
<td>4</td>
<td>3.39</td>
<td>ADAM8</td>
<td>0.0105</td>
<td>1</td>
<td>1.15</td>
</tr>
<tr>
<td>PXDN</td>
<td>0.0139</td>
<td>0</td>
<td>0.00</td>
<td>MAGEA10</td>
<td>0.0105</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>KCNN5</td>
<td>0.0138</td>
<td>0</td>
<td>0.00</td>
<td>F12</td>
<td>0.0105</td>
<td>2</td>
<td>2.02</td>
</tr>
<tr>
<td>DBN1</td>
<td>0.0136</td>
<td>1</td>
<td>1.15</td>
<td>ABCB7</td>
<td>0.0104</td>
<td>1</td>
<td>1.54</td>
</tr>
</tbody>
</table>

NOTE: To avoid biases in representativeness, genes investigated only by HG-U133 Plus 2.0 arrays were excluded from calculations.

Abbreviations: GCS, global node criticality score; #fn, number of first neighbors shared by at least 4 data set; %TN, percentage of shared neighbors with regard to those occurred at least once.
sets when the critical-gene list including FAM53B, TMPO, KIF21B, and WHSC1 were used (Fig. 2B–D; HRs reported in Supplementary Table S4). This model could not be reduced any further without losing prognostic power.

A neighbor-gene model independently predicts overall survival

If the top 100 critical genes indeed represent crucial nodes in the myeloma transcriptional networks, then it could be expected that their interactions, characteristic of the disease, with neighboring genes might be conserved and could also define a set of genes able to predict OS in MM. On the basis of this hypothesis, we focused on genes within the top 100 critical genes that shared a subset of primary neighbors among the different data sets. Setting a threshold of at least 10 conserved primary neighbors in at least 4 series, we identified CCND2, AIF1 (encoding a cytokine-dependent inflammation-responsive protein), and BLNK, encoding

![Graphs showing Kaplan-Meier estimated curves for GSE2658 (A), GSE9782 (B), and GSE15695 (C) cohorts. The samples have been stratified according to the expression profiles of the top 100 critical genes corresponding to 219 probe sets.](image1)

![Graphs showing the Prognostic impact of the critical-gene model in the 3 MM cohorts. A, P values of the 25 probe sets most correlated with OS (P < 0.05); B–D, Kaplan–Meier curves of the 2 groups defined by the critical-gene model (top panels); boxplot of the distribution of the expression values of the 4 genes included in the model (bottom panels) are shown for each data set.](image2)
identified 14 shared neighbors each for the B-cell linker protein (Table 2). In particular, we identified and clinical variables, a Cox proportional hazards model constructed for any probe set using the training set samples, indicating sign [green: positive (POS); red: negative (NEG)]; and significance (P values on y-axis) of correlation with survival. C–E, Kaplan–Meier curves of the 2 groups defined by the neighbor-gene model (top panels); boxplot of the distribution of the expression values of the 4 genes included in the model (bottom panels) are shown for each data set.

Figure 3. Prognostic impact of the neighbor-gene model in the 3 MM cohorts. A, standardized expression profiles in GSE13591 data set of BLNK and the 11 BLNK first neighbor shared in at least 4 networks are shown, ordered according to increasing BLNK expression level. B, proportional hazards models constructed for any probe set using the training set samples, indicating sign [green: positive (POS); red: negative (NEG)]; and significance (P values on y-axis) of correlation with survival. C–E, Kaplan–Meier curves of the 2 groups defined by the neighbor-gene model (top panels); boxplot of the distribution of the expression values of the 4 genes included in the model (bottom panels) are shown for each data set.

Multivariate analyses confirmed independency of the 2 predicting models

To assess the independence of the gene signatures identified and clinical variables, a Cox proportional hazards...
A regression model was fitted by using each gene signature and a number of clinical variables (i.e., serum β₂-microglobulin and albumin, del(13q), t(4;14), t(11;14), sex, and age) in GSE9782 and GSE15695 series. Both the critical- and the neighbor-gene models added predictive power to clinical variables (Table 3). A Cox proportional hazards model was also used to show that both the critical-gene and the neighbor-gene signatures retain independent prognostic effects and add prognostic value to therapy (TT2 and TT3) and to the prediction of outcome in the 7 prognostic groups (3) in GSE2658 (critical-gene model: \( P = 5.1 \times 10^{-7} \) with treatment and \( P = 3.0 \times 10^{-3} \) with 7 groups classification; neighbor-gene model: \( P = 3.0 \times 10^{-4} \) with treatment and \( P = 8.0 \times 10^{-3} \) with 7 groups classification). Finally, a multivariate analysis showed in GSE15695 the independence of the 2 minimal signatures described here from all the other previously described gene risk models, namely, the UAMS 17-gene signature described by Shaughnessy and colleagues (5), the IFM 15-gene signature by Decaux and colleagues (4) and the Myeloma IX 6-gene signature by Dickens and colleagues (ref. 6; Table 4). The neighbor-gene model retained independence in GSE9782 data set, and the critical-gene model in GSE2658. The independence of the critical- and the neighbor-gene signature was also retained in multivariate analysis in GSE15695 if the 70-gene model from UAMS was used instead of the 17-gene model (Supplementary Table S5).

**Discussion**

In this work, we have reconstructed gene regulatory networks from whole transcriptional profiles of 1,883 plasma cell dyscrasia samples and have identified critical nodal genes with prognostic relevance. Notably, networks and signatures have been derived from virtually all of Affymetrix gene expression data sets of primary myeloma tumors available in public repositories. The integration of multiple, independently generated data allowed us to overcome the limited number of samples in individual studies, led to more reliable results, and, ultimately, provided disease biomarkers that seem to be population independent and may be of global relevance. Reconstructing regulatory network from expression data, in this fashion, is a crucial step to understand the mechanisms underlying biological systems. Among several algorithms, ARACNe has been shown

**Table 3.** Multivariate analysis comparing both the critical- and the neighbor-gene models with clinical variables in GSE9782 and GSE15695 data sets

<table>
<thead>
<tr>
<th>Factor</th>
<th>GSE9782 Critical gene</th>
<th>GSE9782 Neighbor gene</th>
<th>GSE15695 Critical gene</th>
<th>GSE15695 Neighbor gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratification on gene model</td>
<td>2.0E-05</td>
<td>0.0014</td>
<td>0.013</td>
<td>0.049</td>
</tr>
<tr>
<td>Serum β₂-microglobulin</td>
<td>0.65</td>
<td>0.12</td>
<td>3.4E-04</td>
<td>9.4E-04</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>3.6E-03</td>
<td>3.7E-04</td>
<td>0.034</td>
<td>0.081</td>
</tr>
<tr>
<td>Del(13q)</td>
<td>nd</td>
<td>nd</td>
<td>0.86</td>
<td>0.5</td>
</tr>
<tr>
<td>t(4;14)</td>
<td>0.029</td>
<td>0.56</td>
<td>0.48</td>
<td>0.6</td>
</tr>
<tr>
<td>t(11;14)</td>
<td>nd</td>
<td>nd</td>
<td>0.14</td>
<td>0.28</td>
</tr>
<tr>
<td>Sex</td>
<td>0.18</td>
<td>0.23</td>
<td>nd</td>
<td>nd</td>
</tr>
<tr>
<td>Age</td>
<td>0.31</td>
<td>0.25</td>
<td>nd</td>
<td>nd</td>
</tr>
</tbody>
</table>

Abbreviation: nd, not determined.

**Table 4.** Multivariate analysis comparing both the critical- and neighbor-gene models with other prognostic gene models in GSE15695, GSE9782, and GSE2658 (TT2)

<table>
<thead>
<tr>
<th>Gene signature</th>
<th>GSE15695 Multivariate Cox regression P value</th>
<th>GSE9782 Multivariate Cox regression P value</th>
<th>GSE2658 (TT2) Multivariate Cox regression P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical-gene model</td>
<td>0.03</td>
<td>0.09</td>
<td>0.018</td>
</tr>
<tr>
<td>Neighbor-gene model</td>
<td>0.011</td>
<td>0.026</td>
<td>0.41</td>
</tr>
<tr>
<td>IFM 15-gene</td>
<td>0.53</td>
<td>0.0061</td>
<td>0.56</td>
</tr>
<tr>
<td>UAMS 17-gene</td>
<td>0.038</td>
<td>7.1E-07</td>
<td>0.00018</td>
</tr>
<tr>
<td>6-gene signature</td>
<td>0.044</td>
<td>0.042</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Abbreviations: IFM, Intergroupe Francophone du Myelome; UAMS, University of Arkansas for Medical Sciences.
to outperform other Bayesian-based methods in its precision for the reconstruction of transcriptional networks (7, 9, 27, 28). To estimate gene–gene correlation, ARACNe uses the MI, which does not assume a linear relationship between the gene expression levels and does not require any assumption about the distribution of variables.

The first aim of this comprehensive analysis was to extract de novo knowledge of the molecular mechanism underlying MM through the investigation of common characteristics of different networks. Network robustness strongly relies on its structure and in particular on the existence of paths between nodes. In this context, an important aim of this analysis was to identify critical components, that is, nodes (genes) or edges (connections) that are important for the functioning of the network. A simple way to assess the importance of these elements is to understand the drop in network stability caused by the deactivation of each individual element at a time (10). Thus, the result of this analysis is based on the idea that the most important genes are not necessarily the most interconnected, but those whose removal causes the largest effect on the network. Of note, in the analysis of the B-cell networks by Basso and colleagues (7), one of the most interesting element was c-MYC, which had only 56 neighbors and ranked 410th in terms of connectivity. However, c-MYC could be considered a critical element as its neighbors could modulate a substantial percentage of all genes in the cell (7). Recently, several procedures have been proposed to further strengthen the ability of this type of analysis to identify critical regulatory elements in reconstructed networks in cancer (11, 12). Thus, we supplemented ARACNe with a method that exploits the topologic features of gene regulatory networks to identify individual components that are biologically relevant, namely, critical genes. The importance of such a procedure depends on the fact that it allows the identification of genes that might play a pivotal role in MM disease irrespective of being deregulated in association with specific molecular features and/or genetic lesions, a condition that might occur but is not mandatory. Moreover, the virtues of this methodology are that it takes advantage of a composite of data sets from multiple institutions and does not select genes on the basis of survival criteria. Specifically, we ranked all the genes of the 7 reconstructed regulatory networks according to their global criticality score. The results of this analysis identified the most critical genes and, interestingly, top among these were identified CCND1 and CCND2, which strongly supported the validity of the approach in MM (29).

Additionally, we identified several other critical genes including WHSC1, FGFR3, RUNX1, MUC1, CD19, CD33, and TNES5FR. Already known as associated with myeloma (2, 30–34). Beyond these genes, other important nodes involving immune response pathways (PLCG2, SYK, BLNK) were also identified, whose involvement in MM is still undefined.

On the basis of this comprehensive analysis of a large panel of transcriptional data, one of our major efforts was directed at the definition of a shared classifier, capable of predicting clinical outcome across the different trials examined. The reverse engineering analysis used in this work represents a solution able to connect apparently discordant information derived from different data sets (3–5). In addition, this approach allows the identification of genes whose detection might be masked by the lack of association with other shared molecular features, thus overcoming limitations imposed by conventional supervised approaches. In this work, we defined 2 models that retained predictive power for OS in multiple, independent data sets: the critical-gene model (including 4 genes, FAM53B, KIF21B, WHSC1, and TMPO) and the neighbor-gene model (including the BLNK primary neighbors, CSGALNACT1 and SCL7A7). In the critical-gene model, we identified the WHSC1 gene, which is deregulated by t(4;14) translocation and is associated with poor prognosis in MM; however, very little was known about the remaining 3 genes in myeloma. For instance, TMPO, is involved in the structural organization of the nuclear envelope, has been found deregulated in several hematologic tumors (35), and is included in the 17-gene signature of Zhan and colleagues (3), whereas KIF21B, encoding a motor protein essential to cellular transport, maps to 1q arm (1q32.1) whose amplification characterizes a significant fraction of high-risk MM tumors.

An intriguing finding of this study is the identification of BLNK, encoding B-cell linker protein, as candidate for a critical role in the outcome of MM. Its normal role in B-cell development is to bridge B-cell receptor–associated kinase activation with downstream signaling pathways. Loss of this protein has been seen in some cases of pre-B acute lymphoblastic leukemia, consistent with a role as a tumor suppressor gene (36). Although BLNK was associated with only a limited number of neighbors in the analyzed networks, several of these were shared among the different data sets: PDCD4 in 7; CDKN1B, MARCKS, TBC1D9, CSGALNACT1 in all but 1 data set; P2RY10 in 5; and DAPPI, C11orf80, CYBB, PNOC, and SCL7A7 in 4 (Fig. 3A; Supplementary Fig. S3). Among these genes the commonly shared neighbor PDCD4 is thought to play a pivotal role in several cancer types, potentially acting as a tumor suppressor and controlling cell proliferation (37). It has also been reported that CDKN1B/p27kip1 interacts with BLNK through JAK3, suppressing pre-B-cell leukemogeneis (38), and plays a role in cell-cycle control in MM (39, 40).

With regard to the neighbor-gene model, CSGALNACT1 has been recently proposed by Bret and colleagues (41) to be associated with good prognosis in MM. Our analysis extends and reinforces their observations. Bret and colleagues only considered the TT2 fraction of GSE2658 data set and compared the highest and the lowest expression quartiles. Here we show that the expression profiles of CSGALNACT1, together with SCL7A7, are capable of stratifying MM patients in 2 groups with different prognosis. CSGALNACT1 encodes a protein involved in glycosaminoglycan chain synthesis and modifications, and it is likely to act in posttranslational modifications downstream of signals activated via SLC family cationic amino acid transporters.
Moreover, the expression profiles of the 2 genes seem to be strictly related, because SLC7A7 is a putative candidate of being part of the pathways, carrying out posttranslational modifications of chondroitin sulfate in MM.

Importantly, we showed that either model retained its prognostic power in comparison with a number of clinical variables and in comparison with other gene-risk models. Specifically, both critical- and neighbor-gene model provide comparable results to UAMS-17 (or UAMS-70) and the 6-gene signature in GSE15695 database. In addition, the critical-gene model retained significance in GSE2658 and the neighbor-gene model in GSE9782, in both of which UAMS-17 (or UAMS-70) signature was largely the most significant. Taken as a whole, these data indicated that the 2 critical- and neighbor-gene models provide additional information to prognostic classification in relationship with the existing ones.

Overall, the analysis of network critical components, based on a large panel of primary tumors, shows that it is possible to reverse engineer transcriptional regulatory networks and unravel crucial nodal points with prognostic relevance. This strategy allowed us to generate 2 robust models with a sensibly reduced number of genes in comparison with others previously reported, both useful to evaluate clinical outcome without suffering any data set–specific bias.

References


Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments

The authors thank Prof. Riccardo Dalla-Favera (University of Columbia, NY) for the critical assessment of the manuscript, and Prof. Massimo Marchiori, Dr. Lino Possamai (both from University of Padua, Italy), and Dr. Adrian Andronache (University of Milan, Italy) for the informatics contributions in the set up of C++ procedures.

Grant Support

This study was supported by research funding from Associazione Italiana Ricerca sul Cancro (AIRC) to A. Neri, University of Modena (Finanziamento Linee Strategiche di Sviluppo dell’Anees, Medicina Molecolare e Rigenerativa, 2008) and Fondazione Cassa di Risparmio di Modena research grant 2007 to S. Bicciato; The Royal Marsden NHS Foundation Trust, NIH Biomedical Research Centre and the NIH Biomedical Research Centre at the Royal Marsden Hospital to B.A. Walker and G.J. Morgan. L. Agnelli was supported by a fellowship from Fondazione Italiana Ricerca sul Cancro (FIRC).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received March 9, 2011; revised June 23, 2011; accepted August 29, 2011; published OnlineFirst September 2, 2011.


Clinical Cancer Research

The Reconstruction of Transcriptional Networks Reveals Critical Genes with Implications for Clinical Outcome of Multiple Myeloma

Luca Agnelli, Mattia Forcato, Francesco Ferrari, et al.

Clin Cancer Res  Published OnlineFirst September 2, 2011.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-11-0596

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2011/09/06/1078-0432.CCR-11-0596.DC1
http://clincancerres.aacrjournals.org/content/suppl/2011/11/22/1078-0432.CCR-11-0596.DC2

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://clincancerres.aacrjournals.org/content/early/2011/11/16/1078-0432.CCR-11-0596.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC)
Rightslink site.