BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 6 2012, pages 823–830
doi:10.1093/bioinformatics/bts059

Gene expression

Advance Access publication January 31, 2012

# Predicting relapse prior to transplantation in chronic myeloid leukemia by integrating expert knowledge and expression data

K. Y. Yeung[1,*], T. A. Gooley[2], A. Zhang[2], A. E. Raftery[3], J. P. Radich[2] and V. G. Oehler[2]

[1]Department of Microbiology, University of Washington, Seattle, WA 98195, [2]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109 and [3]Department of Statistics, University of Washington, Seattle, WA 98195, USA

Associate Editor: David Rocke

**ABSTRACT**

**Motivation:** Selecting a small number of signature genes for accurate classification of samples is essential for the development of diagnostic tests. However, many genes are highly correlated in gene expression data, and hence, many possible sets of genes are potential classifiers. Because treatment outcomes are poor in advanced chronic myeloid leukemia (CML), we hypothesized that expression of classifiers of advanced phase CML when detected in early CML [chronic phase (CP) CML], correlates with subsequent poorer therapeutic outcome.

**Results:** We developed a method that integrates gene expression data with expert knowledge and predicted functional relationships using iterative Bayesian model averaging. Applying our integrated method to CML, we identified small sets of signature genes that are highly predictive of disease phases and that are more robust and stable than using expression data alone. The accuracy of our algorithm was evaluated using cross-validation on the gene expression data. We then tested the hypothesis that gene sets associated with advanced phase CML would predict relapse after allogeneic transplantation in 176 independent CP CML cases. Our gene signatures of advanced phase CML are predictive of relapse even after adjustment for known risk factors associated with transplant outcomes.

**Availability:** The source codes and data sets used are available from the web site http://expression.washington.edu/publications/kayee/integratedBMA.

**Contact:** kayee@u.washington.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on June 14, 2011; revised on January 24, 2012; accepted on January 25, 2012

## 1 INTRODUCTION

The prediction of the diagnostic category of a tissue sample from its expression array phenotype given the availability of similar data from tissues in identified categories is known as 'classification' (or 'supervised learning'). In the context of gene expression data, the samples are usually the experiments, and the classes are usually different types of tissue samples, for example, cancer versus non-cancer e.g. (Alon *et al.*, 1999), different tumor types e.g. (Ramaswamy *et al.*, 2001), prognostic outcomes e.g. (van 't Veer *et al.*, 2002), or different stages of disease e.g. (Radich *et al.*, 2006). A challenge in predicting the diagnostic categories using gene expression data is that the number of genes is usually much greater than the number of tissue samples available, and only a subset of the genes is relevant in distinguishing different classes. Selection of predictive signature genes for classification is known as 'variable selection' or 'feature selection'. A wide variety of algorithms have been proposed in the literature to select predictive signature genes, e.g. (Abeel *et al.*, 2010; Chen *et al.*, 2009; Lee *et al.*, 2003; Nguyen and Rocke, 2002; Tusher *et al.*, 2001; Yang and Song, 2010). A small set of relevant genes is essential for the development of inexpensive diagnostic tests.

Different feature selection algorithms can potentially select different relevant genes, different numbers of relevant genes and lead to different classification accuracy (Xu and Wong, 2010). Ein-Dor *et al.* (2005) reported that gene selection is heavily influenced by the subset of patients even when the feature selection method and data set stayed constant. This is mainly due to the fact that many genes have similar correlations with the class labels, and much larger training sets are needed to generate a robust gene list (Ein-Dor *et al.*, 2006). The lack of overlap between the genes of different prognostic signatures is well documented, e.g. (Drier and Domany, 2011; Dupuy and Simon, 2007; Michiels *et al.*, 2005). Ioannidis (2005) commented that the discovery of the true gene signature remains a challenge, and emphasized the importance of independent validation and large sample size. Another critique of existing feature selection methods includes the lack of biological meaning of the signature genes (Drier and Domany, 2011).

In this article, we present a novel approach that integrates expert knowledge and predicted functional relationships with gene expression analysis to derive gene signatures. We aim to select 'robust' signature genes that are 'both predictive and biologically relevant' to the mechanism underlying the disease of interest. Our method uses expert knowledge and predicted functional relationships to guide our search for signature genes among the many possible genes that are highly correlated with the class labels and with each other. Our method is based on the idea that genes that are functionally related to a reference gene set known to be associated with the disease of interest are more likely to be biologically relevant to this disease. Figure 1 shows an overview of our method.

### 1.1 Case study: chronic myeloid leukemia progression

Chronic myeloid leukemia (CML) serves as our case study in this work. CML typically presents in chronic phase (CP) and if not

*To whom correspondence should be addressed.

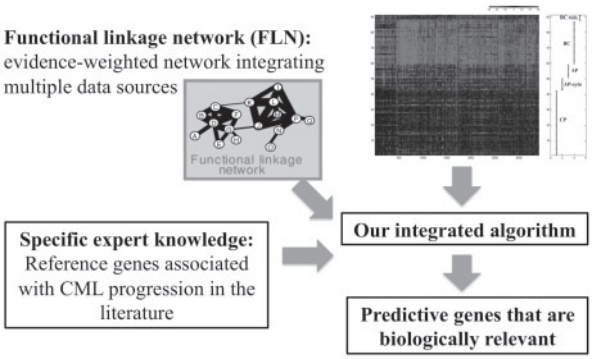


**Fig. 1.** Overview of our method integrating expert knowledge, predicted functional relationships and microarray data to derive predictive genes that are biologically relevant to the disease of interest.

treated effectively progresses through accelerated phase (AP) to the acute leukemia phase of blast crisis (BC). A reciprocal translocation between chromosomes 9 and 22 yields the Bcr-Abl fusion protein that is the primary driver of CML disease. Tyrosine kinase inhibitors (TKIs), such as imatinib (IM), dasatinib and nilotinib, which target the constitutively active Abl tyrosine kinase, are effective first-line therapy for CP CML (Druker *et al.*, 2001). However, TKIs are significantly less effective in AP or BC CML where not only Bcr-Abl dependent, but also Bcr-Abl independent mechanisms are at work (Druker *et al.*, 2006; Faderl *et al.*, 1999; Hansen *et al.*, 1998; Melo and Barnes, 2007; Quintas-Cardama and Cortes, 2009; Ren, 2005; Sawyers *et al.*, 2002; Shah, 2008). Allogeneic hematopoietic stem cell transplantation (HSCT) is generally reserved for CP patients who are resistant to TKI therapy, or patients who have AP or BC CML. HSCT is associated with >80% survival when performed in CP, but outcomes are significantly worse for advanced disease, ~40% in AP and <20% in BC (Hansen *et al.*, 1998; Radich *et al.*, 2003).

There are no established molecular predictors of transplant outcome in CML, thus clinical measures such as the Sokal, Hasford, or European Group for Blood and Marrow Transplantation (EBMT) Risk Scores have been used for prognostication (Gratwohl *et al.*, 1998; Hasford *et al.*, 1996; McWeeney *et al.*, 2010; Mohty *et al.*, 2007; Sokal *et al.*, 1984; Yong *et al.*, 2006). The EBMT score has been validated extensively, and is used to predict outcomes prior to HSCT (Gratwohl *et al.*, 1998; Passweg *et al.*, 2004). The variables that comprise the EBMT score are: phase of disease (CP, AP, or BC), transplant donor type (related donor versus unrelated donor), donor/recipient sex combination (female donor into male recipient versus all other combinations), patient age and interval from diagnosis to transplantation (Gratwohl *et al.*, 1998). However, the EBMT risk score does not entirely account for the heterogeneity observed in transplant outcomes. Consequently, our goal is to identify molecular prognostic predictors to supplement known clinical risk variables and guide risk-adjusted therapy at diagnosis.

We previously applied a probabilistic method called iterative Bayesian model averaging (iBMA) to a microarray data set consisting of patients in different phases of CML (Oehler *et al.*, 2009b) and identified a 6-gene signature that accurately discriminated CP from BC. BMA is a multivariate method that accounts for the uncertainty in the selection of signature genes by averaging over multiple models (Hoeting *et al.*, 1999; Raftery, 1995), i.e. sets of potentially overlapping predictive genes. It yields posterior probabilities of the predictions and of the inclusion of each gene in the model. We extended BMA to be applied to high dimensional gene expression data and showed that the iBMA algorithm identifies small gene sets with high prediction accuracy (Yeung *et al.*, 2005). We also showed that our 6-gene signature discriminated CP, AP and BC patients and hypothesized a signature associated with progression when detected in CP may predict poorer outcomes, because expression of these genes reflects evidence of more advanced disease at the molecular level in CP patients (Oehler *et al.*, 2009b). In this study we build upon our iBMA method and validate this hypothesis.

### 1.2 Our contributions

We present a novel integrated method that by incorporating genes relevant to CML disease and progression not only produced accurate class predictions, but also yielded gene signatures that are more robust and stable than using gene expression data alone. Our method can be applied to other high-throughput data, including the expression of genes, proteins, microRNAs, or single nucleotide polymorphism variations. Our findings were substantiated in 196 independent CML patient samples. Specifically, we show that a molecular signature associated with disease progression when detected in CP patients prior to HSCT drives outcomes after transplantation even after adjusting for contributions from known clinical risk factors.

## 2 METHODS

### 2.1 BMA

BMA takes model uncertainty into consideration by averaging over the posterior distributions of a quantity of interest based on multiple models, weighted by their posterior model probabilities (Hoeting *et al.*, 1999; Raftery, 1995). In the case of binary classification, let $Y$ be the class of a sample in the test set, where $Y=0$ or 1, and let $D$ be the 'training data set' for which the classes are known. In BMA, the posterior probability of $Y=1$ given the training set $D$ is the weighted average of the posterior probability of $Y=1$ given the training set $D$ and model $M_k$ multiplied by the posterior probability of model $M_k$ given training set $D$, summing over a set of models $M_k$ for $k$ in $B$, where $B$ is a set of indices, mathematically, $\Pr(Y=1|D)=\sum_{k\in B}\Pr(Y=1|D,M_k)*\Pr(M_k|D)$. We used logistic regression (Hosmer and Lemeshow, 2000) to evaluate $\Pr(Y=1|D, M_k)$. When classifying samples on microarray data, our goal is to identify the relevant genes (or variables). The posterior probability of gene $x_i$ is equal to the sum of the posterior probabilities of all selected models $M_k$ that include this gene. Hence, all relevant genes are included in at least one chosen model.

In order to efficiently identify a reduced set of models $M_k$ for the weighted average calculations, Raftery (1995) used the leaps and bounds algorithm (Furnival and Wilson, 1974) which rapidly returns the best *nbest* models of each size up to $w$ genes ($w=30$, $nbest=10$ in our experiments). Madigan and Raftery (1994) proposed the Occam's window method as a way of choosing a set of parsimonious and data-supported models. Their idea is to discard models that are much less likely than the best model supported by the data (the default is 20 times less likely in terms of model likelihood). Therefore, the set of models used in the weighted average calculations is chosen by first applying the leaps and bounds algorithm, and then the Occam's window method. We used the Bayesian information criterion (BIC) to approximate the posterior probability of a model $M_k$ (Kass and Raftery, 1995). In this study, we used the R package 'BMA' and the bioconductor package 'iterativeBMA'.

## 2.2 iBMA algorithm

To apply BMA to the high-dimensional gene expression data, we used the iterative BMA (iBMA) method of Yeung *et al.* (2005). In iBMA, we first ranked the genes in order with a univariate gene selection method and then successively applied BMA to the ordered genes. In the univariate ranking step, genes with relatively large variation between classes and relatively small variation within classes received high rankings. We then applied BMA to the $w$ top ranked genes ($w=30$). Then variables to which the BMA algorithm assigned low posterior probabilities (<5%) of being in the predictive model were removed. Suppose $m$ variables were removed. The next $m$ variables from the rank ordered $R^2$ were added so that we maintained a window of $w$ variables and applied BMA again. These steps of gene swaps and iterative applications of BMA were continued until we considered all top $v$ variables in our univariate ranked gene list.

## 2.3 Selection of reference genes

Reference genes represented input expert knowledge that are chosen independent of the gene expression data. They were compiled from pathways involved in CML disease and CML disease progression and included genes involved in Bcr-Abl signaling, the stem cell associated pathways Hedgehog and WNT, as well as individual candidates known to be associated with disease progression (e.g. SET, PRAME) or have been described in leukemia stem/progenitor cells (Jamieson *et al.*, 2004; Melo and Barnes, 2007; Neviani *et al.*, 2005; Oehler *et al.*, 2009a; Quintas-Cardama and Cortes, 2009; Stirewalt *et al.*, 2008; Zhao *et al.*, 2009) (Supplementary Table S1). An expanded set included all genes in the base set, but included additional candidates from the pathways and sources listed above (Supplementary Table S2).

## 2.4 Functional linkage network

The functional linkage network (FLN) is an evidence-weighted network aimed to represent the functional associations between human genes related to different diseases (Linghu *et al.*, 2009). The FLN was constructed using a naïve Bayes classifier integrating various functional genomics features, including co-expression data, protein–protein interactions, protein domain sharing, co-occurrence in PubMed abstracts and functional associations mapped from other organisms. The nodes in FLN represented individual genes and the weighted undirected edges (quantified by the log likelihood ratio scores, LR) represented the degree of their overall functional association upon combining various data sources. The FLN is extremely dense, with the average number of neighbors per gene approximately 2000 even with a LR score cut-off at 1. Linghu *et al.* (2009) illustrated the utility of the FLN in the prioritization of candidate genes in various diseases, and proposed that the associations identified by the FLN could be used to derive novel hypotheses on molecular mechanisms underlying diseases and therapies.

## 2.5 Integrated iBMA algorithm

The FLN is a weighted undirected graph consisting of 21 657 nodes and ~22.4 million edges with a median edge weight (LR score) of 0.21 (Linghu *et al.*, 2009). The nodes (genes) of the FLN are represented by Entrez Gene IDs. We obtained a sub-graph of the full FLN consisting of genes profiled in the CML progression microarray data (see Section 2.7). This sub-graph (denoted as FLN-sub) consisted of 11 514 nodes and ~11.7 million edges. As we increased the LR score threshold from log(5) to log(50), the average degree per node in FLN-sub reduced from 31.4 to 4.2. We empirically determined a LR score threshold for the edges by evaluating the average prediction accuracy over 3-fold cross-validation repeated 100 times using the gene expression data (see Table 1). Supplementary Figure S1 shows a flow chart of this process.

After mapping the reference genes to their unique Entrez Gene IDs, we located them on the FLN and identified neighbor genes that were connected to these reference genes (i.e. path length 1 from the reference genes). We then took a union of the genes in the reference and the neighbor sets, and applied iBMA (see Fig. 2). Table 1 shows the number of genes after the union at various thresholds. The integrated iBMA approach is outlined and studied in detail in Supplementary Methods.

A pictorial view of our integrated iBMA algorithm is shown in Figure 2: (1) Remove edges in FLN with a relatively small weight; (2) Locate the reference genes known to be associated with CML progression on the FLN; (3) Identify neighbor genes connected to the given reference genes in the FLN (i.e. these neighbor genes are functionally related to the reference genes);nd (4) Apply iBMA to the union of the reference and neighbor sets.

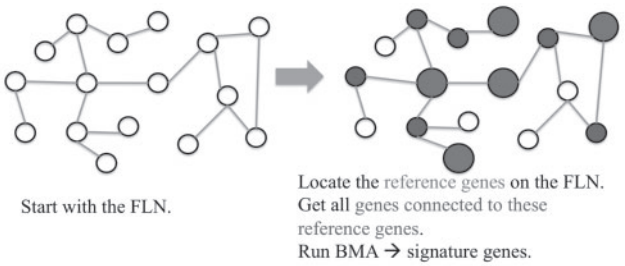


**Fig. 2.** Illustration of our integrated iterative BMA algorithm. The reference genes are shown as large red nodes, and the genes connected to the reference genes are shown as small blue nodes.

**Table 1.** Prediction accuracy of signature genes on the microarray data using the base reference set consisting of 27 genes

| Edge threshold | # genes in UNION (reference + neighbor) | Average # genes | Average Brier score | Average % errors |
|---|---|---|---|---|
| log(5) | 2927 | 10.56 | 0.62 | 0.85 |
| log(10) | 1377 | 10.04 | 0.43 | 0.26 |
| log(20) | 764 | 9.63 | 0.49 | 0.17 |
| log(50) | 421 | 9.34 | 0.68 | 0.53 |
| No FLN | 12 734 | 12.93 | 0.54 | 0.64 |

The average Brier Score and percentage of errors are obtained from 3-fold cross-validation, repeated 100 times.

## 2.6 Microarray data studying CML progression

We used the CP (42) and BC (30) patient samples in the published CML progression microarray data (Radich *et al.*, 2006), which is available from the Gene Expression Omnibus (GEO) database with accession number GSE4170. After filtering out probes on the arrays that were not mapped to any Entrez Gene IDs, 12 734 genes were used in our analyses.

## 2.7 Evaluation of predictive performance on microarray data

We evaluated predictive performance on the microarray data using 3-fold cross-validation repeated 100 times. We computed the average percentage error as the average fraction of classification errors over 72 samples (42 CP + 30 BC) in 100 cross-validation runs. We also adopted the average Brier Score (Brier, 1950) that accounts for the magnitudes of predicted probabilities. The Brier Score is defined as the sum of squares of the difference between the true class label and the predicted probability over all samples. If the predicted probabilities are constrained to equal to 0 or 1, the Brier Score is equal to the total number of classification errors.

### 2.8 Quantitative PCR validation data

Quantitative PCR (qPCR) was used to validate gene expression in 222 patient samples (176 CP, 23 AP, 14 BC-remission and 9 BC). Bone marrow samples were obtained from Institutional Review Board approved protocols. Written informed consent is according to the Declaration of Helsinki. Among CP patients, 26 patients had samples assessed by both qPCR and microarray gene expression (Radich *et al.*, 2006), whereas 196 samples were uniquely assessed by qPCR. Gene expression, performed in duplicate, was quantitated and normalized to GUSB expression (control gene) by qPCR (Applied Biosystems) as previously described (Oehler *et al.*, 2009b; Radich *et al.*, 1995, 2001).

### 2.9 Analyses of qPCR data

We divided our CP patients further into CP-early (120 patients) and CP-late (56 patients). Patients defined as CP-early were within 1 year of diagnosis at the time when the sample was drawn. Patients with early CP are less likely to develop resistance on TKI therapy (Branford *et al.*, 2003) or relapse after allogeneic transplantation (Gratwohl *et al.*, 1998; Passweg *et al.*, 2004). We transformed our qPCR data logarithmically and ignored patient samples with missing data. Due to the limited number of BC patient samples ($n=9$), we computed prediction models for CP-early versus AP (120 CP-early versus 23 AP samples). Since the microarray data and the qPCR data are on different scales, we re-fitted the BMA models using cross-validation. In other words, after we identified the signature genes by applying our integrated iBMA method to the CML progression microarray data, we computed the regression coefficients of these signature genes using 10-fold cross-validation on the qPCR data.

We used 'bagging' to address the issue of the unbalanced class sizes (120 CP-early versus 23 AP samples). In each fold of the cross-validation, let Train-large be the training data of the larger class (i.e. CP-early) and let Train-small be the training data of the smaller class (i.e. AP). Let $n$-large and $n$-small be the sizes of the training data Train-large and Train-small, respectively. Also, let the test data of the larger and smaller classes be Test-large and Test-small, respectively. We then repeated the following bagging step 20 times: we randomly sampled $n$-small samples in Train-large with replacement, computed the BMA models using these $n$-small random samples from Train-large and all the samples from Train-small, and predicted the test samples in Test-large and Test-small using this model. The predicted class label of each test sample is equal to the majority class over these 20 bagging runs. In other words, each of these 20 bagging runs carried a weight of 1/20, and the predicted probability for each test sample is equal to the number of times this sample is predicted to be class 1 divided by 20. We observed that different numbers of bagging runs produced similar results (data not shown).

### 2.10 Patient and transplant characteristics

Patient and transplant characteristics and calculation of the EBMT score for 176 CP CML patients are shown in Supplementary Table S3. The median sample draw date prior to transplantation was 22 days prior (range, 0–250 days prior). Transplants occurred between 1993 and 2007. The majority of patients did not receive TKI prior to transplantation. IM was approved by the FDA as second-line therapy in 2001 and as first-line therapy in 2003. Fifty-one patients underwent allogeneic transplantation from 2002 onwards and improved cytogenetic responses prior to transplantation suggest these patients were treated with TKI prior to transplantation. Cytogenetic responses were characterized as follows: complete [0% Philadelphia (Ph)-positive cells], partial (1–34% Ph-positive cells), minor (35–65% Ph-positive cells), minimal (66–95% Ph-positive cells), or no (96–100%). Kaplan–Meier estimates were used to summarize the probability of overall and disease-free survival. Cumulative incidence estimates were used to summarize the probability of relapse and non-relapse mortality (NRM). Relapse was regarded as a competing risk for NRM, and death without relapse a competing risk for relapse. Cox regression was used to assess the association of the iBMA predicted probability with each of these time-to-event outcomes, and the iBMA predicted probability was modeled as a continuous linear variable. All models were adjusted for EBMT Risk Score. The impact of total white blood cell count, blast percent and cytogenetic response on outcomes was also modeled as a continuous variable.

## 3 RESULTS

In this work, we present a novel integrated iBMA algorithm that identifies robust and biologically relevant predictors by integrating prior expert knowledge, predicted functional relationships and gene expression data. Prior expert knowledge is represented as a set of reference genes that are selected independently of the gene expression data, and consists of genes known to be associated with the disease of interest in the literature (see Section 2.3 for details). We make use of predicted functional relationships represented in the FLN (Linghu *et al.*, 2009). The FLN is an evidence-weighted network (i.e. a weighted undirected graph) computed from various functional genomics data sources and aimed to represent the functional associations between human genes related to different diseases, and has been used to prioritize candidate genes in various diseases (Linghu *et al.*, 2009). The FLN represents universal prior knowledge that can be used in any disease of interest.

### 3.1 Assessment using the microarray data

*3.1.1 Prediction accuracy* We applied integrated iBMA to classify CP versus BC patient samples in the CML progression microarray data. We experimented with two reference gene sets, the base and expanded set, consisting of 27 and 72 reference genes, respectively (Supplementary Tables S1 and S2), and different edge weight thresholds. We evaluated the prediction accuracy using 3-fold cross-validation repeated 100 times on the microarray data. Table 1 shows the average prediction accuracy using the base reference set consisting of 27 genes. See Supplementary Table S4 for the variance over cross-validation runs. As expected, the average degree in FLN-sub and the number of neighbor genes connected to the reference set decreased as we increased the edge threshold. As the number of neighbor genes is decreased, the highly predictive genes from lower edge thresholds might disappear, and hence, the prediction accuracy might be reduced. As a baseline, applying iBMA to the microarray gene expression data (without the FLN and without the reference genes) selected an average of 12.93 genes, produced an average Brier Score of 0.54 and an average percentage errors of 0.64%. Therefore, the integrated iBMA method using the FLN and reference genes produced similar prediction accuracy while selecting fewer genes on an average.

The edge threshold log(10) produced the lowest average Brier Score from Table 1. Using an edge threshold of log(10), we then applied integrated iBMA to the full CML progression microarray data (using all 72 CP and BC patient samples) and selected six signature genes 'base-log10' = (RALGDS, LASP1, G6PD, ADRBK1, LRPPRC, PSMA1) with posterior probabilities >5%. RALGDS is represented by more than one probe on the array platform, and two splice variants (AB037729 and AK000242) are selected with posterior probabilities 22.3 and 15.1%, respectively. The remaining five signature genes LASP1, G6PD, ADRBK1, LRPPRC, PSMA1 are selected with posterior probabilities 16.6, 15.1, 15.1, 13.7 and 13.7%, respectively. Figure 3 shows how these six signature genes are connected to our reference genes in the FLN at this threshold. Specifically, SET, SMO, SHH, RRAS2, ABL1,

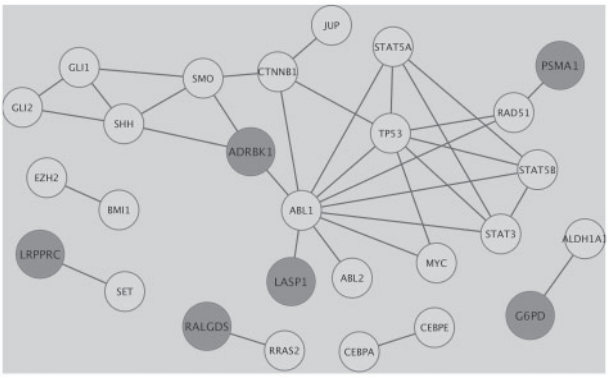


**Fig. 3.** Signature genes (shown as large orange nodes) selected using the base reference genes (shown as small pink nodes) and weight threshold log(10) using the full CML progression microarray data.

RAD51, ALDH1A1 are in the base reference set that are also neighbors of our signature genes. Since the reference set represents prior knowledge (i.e. genes known to be associated with CML from the literature), they might not be predictive in the expression data. Hence, the reference genes are not necessarily chosen by integrated iBMA. SET is over-expressed in BC CML and participates in the inhibition of differentiation seen with progression to the acute leukemic phase of CML (Neviani *et al.*, 2005). SMO and SHH are members of the sonic hedgehog pathway, which is important in maintaining leukemia stem cells in CML (Zhao *et al.*, 2009), whereas ALDH1A1is differentially expressed in acute myeloid leukemia progenitor cells as compared to normal hematopoietic stem cells (Stirewalt *et al.*, 2008). Lastly, RRAS2, ABL1 and RAD51 are associated with Bcr-Abl signaling (Melo and Barnes, 2007; Quintas-Cardama and Cortes, 2009).

The edge threshold log(20) produced similar average Brier Score (0.49) and even lower average percentage error (0.17%) as the log(10) threshold, see Supplementary Figure S2. We also experimented with an expanded reference set consisting of 72 genes, and the prediction accuracy is summarized in Supplementary Table S5. With this expanded reference set, the numbers of neighbor genes are higher than is the case with the base reference genes. The log(20) and log(50) thresholds both yielded better average prediction accuracy than iBMA without the FLN and without the reference genes (average Brier Scores of 0.49 and 0.47, both $<0.54$), see Supplementary Figures S3 and S4.

*3.1.2 Stability of signature genes* We showed that the integrated iBMA algorithm using expert knowledge and functional relationships encoded in the FLN selected more stable signature genes on an average over cross-validation than iBMA using the microarray data alone. Since the numbers of genes selected vary over different folds in cross-validation runs, we defined 'stable' signature genes as frequently selected genes over cross-validation (i.e. gene that are selected in at least 50 or 90% in all folds in cross-validation runs), and quantified stability by the fraction of 'stable' signature genes in each fold and each run.

Using the base reference set and a weight threshold of log(10), the average fractions of signature genes selected in at least 50 and 90% of all cross-validation runs are 0.53 and 0.29, respectively. This is in contrast to iBMA using microarray data alone in which the average fraction of signature genes selected in at least 50 and 90% of all runs are 0.24 and 0, respectively. In other words, without the expert knowledge and FLN, iBMA did not select any signature genes with over 90% stability in any of the 100 cross-validation runs. However, on an average, almost 30% of signature genes selected by integrated iBMA are highly stable, i.e. selected in at least 90% of cross-validation runs. Supplementary Figure S5 compared the fraction of signature genes that are selected in at least 50% of the time in each cross-validation runs using the base reference set and log(10) threshold. There is no clear pattern of stability over edge weight thresholds from Table 2. However, it is clear that integrated iBMA produced more stable signature genes than without using any expert knowledge and no FLN. Note that stability over cross-validation does not necessarily imply stability over different reference genes or FLN edge thresholds.

**Table 2.** Stability of signature genes over cross-validation on the microarray data using the base reference set consisting of 27 genes

| Edge threshold | Average fraction of genes selected in $\geq 50\%$ of cross-validation runs | Average fraction of genes selected in at selected in $\geq 90\%$ of cross |
|---|---|---|
| No FLN | 0.24 | 0 |
| log(5) | 0.36 | 0 |
| log(10) | 0.53 | 0.29 |
| log(20) | 0.37 | 0.20 |
| log(50) | 0.39 | 0.32 |

The average fractions of genes are computed over 3-folds in 100 cross-validation runs.

## 3.2 Analyses of independent qPCR data

To substantiate the predictive power of our signature genes, we profiled a total of 35 signature genes using qPCR in 222 patient samples. The 35 signature genes were derived from applying integrated iBMA to the full CML progression microarray data at various FLN edge thresholds using both sets of reference genes. We also profiled the six signature genes identified from our previous study (Oehler *et al.*, 2009b) using iBMA and microarray data alone, and we denote this gene set as previous6 = (NOB1, DDX47, CD101, LTB4R, SCARB1, SLC25A3). See Supplementary Methods for the details of these 35 genes.

Most patients present in CP and far fewer patients in AP and even fewer patients in BC. This represents the clinical dilemma of not having large (or equal) numbers of patient samples for these studies. Our previous analyses showed that prediction accuracy was generally improved when we developed prediction models for disease phases that were more distinct (Oehler *et al.*, 2009b). Therefore, as described in Methods, we assigned a patient sample to CP-early or CP-late based on time from diagnosis (Gratwohl *et al.*, 1998). Due to the small number of BC patient samples ($n = 9$), the prediction accuracy of CP-early versus BC is not very reliable. Therefore, we focused on CP-early versus AP (120 versus 23 samples). We addressed the issue of the unbalanced class sizes using bagging (Mitchell, 1997) (see Section 2.10 in Methods). Since the qPCR data are on different scales than the microarray gene expression data, we re-fitted the BMA model using cross-validation.

Supplementary Table S6 showed the prediction accuracy of various sets of signature genes classifying CP-early versus AP

patient samples on the qPCR data. We observed that signature gene sets determined using integrated iBMA produced comparable prediction accuracy to the signature gene set determined using microarray data alone (previous6). In fact, the signature genes in expanded-log50 = (RALGDS, FKBP1A, PRKCD, AVEN, PXN, RAC2, FGR) performed the best with average accuracy ~81%. As our signature genes were selected from classifying CP versus BC patient samples on the microarray data, we did not expect prediction accuracies as high as those shown in Table 1 when comparing CP-early to AP.

### 3.3 Gene set expression in 172 CP CML patients prior to HSCT predicts relapse after transplantation

Having identified signature gene sets that accurately classified CP from advanced phase disease, we then tested the hypothesis that a gene signature associated with progression when detected in CP predicts poorer outcome. Such validation would be of significant clinical relevance. Specifically, we studied the association between the predicted probabilities computed from integrated iBMA using each signature gene set and the clinical outcomes (relapse) for 172 CP patients. Among these patients, 45 died and 24 relapsed by last contact. Estimates of overall survival (1- and 5-year) were 85 and 78%, respectively, and 1 and 5-year estimates of relapse were 7 and 12%, respectively.

After adjustment for the EBMT Risk Score, which accounts for known clinical variables associated with outcomes, we found the predicted probabilities computed using integrated iBMA and expression of base-log10 = (RALGDS, LASP1, G6PD, ADRBK1, LRPPRC, PSMA1) when detected prior to allogeneic transplantation had the strongest correlation with subsequent relapse after allogeneic transplantation (see Fig. 4A). In CP patients, an increase of 0.2 in the predicted probabilities was associated with an increase in relapse of 46% [HR = 1.46 (1.06–2.02, $P = 0.02$)]. Expression of the previous6 gene set also demonstrated a trend towards increased relapse after allogeneic transplantation. An increase in predicted probability of 0.2 was associated with an increase in relapse of 23% [HR = 1.23 (0.97–1.57, $P = 0.09$)] (see Fig. 4B). Notably, after adjusting for the 'previous6' signature, an increase in the 'base-log10' signature was still associated with an increase in the risk of relapse [HR = 1.38 (0.99–1.91, $P = 0.05$) for an increase of 0.2 in the iBMA predicted probability]. On the other hand, after adjusting for the 'base-log10' signature, an increase in the predicted probability of the 'previous6' although still associated with an increased risk of relapse, was not statistically significant and the magnitude of the association was actually reduced [HR = 1.17 (0.91–1.50, $P = 0.23$) for each increase in 0.2 of predicted probability]. Neither gene signature was significantly associated with NRM, in keeping with our original hypothesis that genes associated with progression drive resistance to therapy rather than predicting toxicities associated with therapy.

In addition to the EBMT Risk Score, we also examined the impact of several other clinical variables on outcomes including total white blood cell count and blast percentage, as well as cytogenetic response at the time of sample draw (Supplementary Methods and Results). None of these additional factors were statistically significantly associated with relapse, and adjustment for these factors did not alter our conclusions, suggesting that these signatures identify heterogeneity at the molecular level, rather than simply reflecting burden of disease.

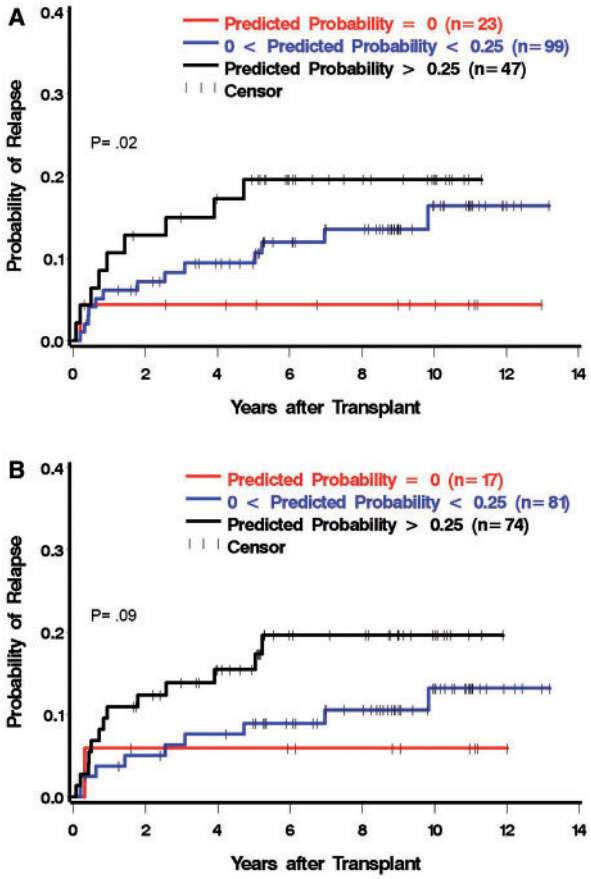


**Fig. 4.** (**A**) After adjustment for the EBMT risk score, the predicted probability of the 'base-log10' signature selected by integrated iBMA correlated with relapse after allogeneic transplantation in CP CML patients. In CP patients, an increase of 0.2 in the probability correlated with an increase in relapse of 46% [HR = 1.46 (1.06–2.02, $P = 0.02$)]. (**B**) After adjustment for the EBMT risk score, we found that increased predicted probability of the 'previous6' gene signature demonstrated a trend towards increased relapse after allogeneic transplantation in CP CML patients. An increase in probability score of 0.2 was associated with an increase in relapse of 23% [HR = 1.23 (0.97–1.57, $P = 0.09$)].

## 4 DISCUSSION

We present a novel method called integrated iBMA that incorporates expert knowledge and predicted functional relationships into expression analyses to identify genes predictive of CML progression. We showed that our method identified signature genes associated with disease progression with high prediction accuracy. We also showed that the signature genes identified by integrated iBMA were relatively stable in cross-validation runs compared to genes selected using microarray data alone. We subsequently profiled selected signature gene sets from independent patient samples using qPCR and showed that our signature genes are predictive of CML phase progression in this independent data set. Lastly, because therapy response is so dependent on CML phase, we hypothesized that advanced phase gene signatures when detected in CP CML would predict poor response. Using integrated iBMA, we derived gene signatures that were independent molecular predictors

of relapse after transplant that withstood adjustment for known clinical variables associated with transplant outcomes.

In regards to the last finding, few molecular predictors have been identified that associate with outcome. BMI1 expression, and more recently the hematopoietic cell transplantation comorbidity index (HCT-CI) together with elevated C-reactive protein (CRP) were found to be associated with inferior survival and increased NRM (Mohty *et al.*, 2008; Pavlu *et al.*, 2010). Our gene signatures, which are associated with relapse, provide complementary information and may better identify which patients require more careful monitoring or early treatment intervention after HSCT. Lastly, because response to all therapies in CML is so dependent on phase, we believe these predictors may also predict outcomes on other therapies such as TKIs.

Our proposed framework can be generalized to other predicted functional relationships such as HEFalMp (Huttenhower *et al.*, 2009) and other variable selection methods such as the lasso (Tibshirani, 1996). Recently, Bessarabova *et al.* (2011) compared the genetic alterations in common human cancers, and identified common signaling pathways and transcription regulation that underlie different genes encoded in these genetic alterations. Our method is similar to their approach in the sense that we adopt the FLN as universal prior knowledge for any disease of interest. Since the expert knowledge is represented as reference genes that are derived from the literature, these reference genes may consist of or interact with the 65 'universal cancer genes' identified by Bessarabova *et al.* Therefore, our method is flexible depending on how specific the inputs are in the form of the reference genes.

## ACKNOWLEDGEMENTS

We thank Dr Roger Bumgarner for valuable discussions.

*Funding*: National Institutes of Health [R01 GM084163 to K.Y.Y., 3R01 GM084163-02S2 to K.Y.Y., R01 CA140371 to V.G.O., R01 HD54511 to A.E.R.]; the Leukemia and Lymphoma Society Translational Research Grant to V.G.O; and the American Society of Hematology Clinical/Translational Scholar Award to V.G.O.

*Conflict of Interest*: none declared.

## REFERENCES

Abeel,T. *et al.* (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, **26**, 392–398.

Alon,U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Bessarabova,M. *et al.* (2011) Functional synergies yet distinct modulators affected by genetic alterations in common human cancers. *Cancer Res.*, **71**, 3471–3481.

Branford,S. *et al.* (2003) Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mutations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis. *Blood*, **102**, 276–283.

Brier,G.W. (1950) Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78**, 1–3.

Chen,P.C. *et al.* (2009) A new regularized least squares support vector regression for gene selection. *BMC Bioinform.*, **10**, 44.

Drier,Y. and Domany,E. (2011) Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes? *PLoS One*, **6**, e17795.

Druker,B.J. *et al.* (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.*, **344**, 1031–1037.

Druker,B.J. *et al.* (2006) Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N. Engl. J. Med.*, **355**, 2408–2417.

Dupuy,A. and Simon,R.M. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl Cancer Inst.*, **99**, 147–157.

Ein-Dor,L. *et al.* (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Ein-Dor,L. *et al.* (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl Acad. Sci. USA*, **103**, 5923–5928.

Faderl,S. *et al.* (1999) Chronic myelogenous leukemia: biology and therapy. *Ann. Intern. Med.*, **131**, 207–219.

Furnival,G.M. and Wilson,R.W. (1974) Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

Gratwohl,A. *et al.* (1998) Risk assessment for patients with chronic myeloid leukaemia before allogeneic blood or marrow transplantation. Chronic Leukemia Working Party of the European Group for Blood and Marrow Transplantation. *Lancet*, **352**, 1087–1092.

Hansen,J.A. *et al.* (1998) Bone marrow transplants from unrelated donors for patients with chronic myeloid leukemia. *N. Engl. J. Med.*, **338**, 962–968.

Hasford,J. *et al.* (1996) Analysis and validation of prognostic factors for CML. German CML Study Group. *Bone Marrow Transplant.*, **17** (Suppl. 3), S49–S54.

Hoeting,J.A. *et al.* (1999) Bayesian model averaging: a tutorial. *Stat. Sci.*, **14**, 382–401.

Hosmer,D.W. and Lemeshow,S. (2000) *Applied Logistic Regression*. Wiley, New York.

Huttenhower,C. *et al.* (2009) Exploring the human genome with functional maps. *Genome Res.*, **19**, 1093–1106.

Ioannidis,J.P. (2005) Microarrays and molecular research: noise discovery? *Lancet*, **365**, 454–455.

Jamieson,C.H. *et al.* (2004) Granulocyte-macrophage progenitors as candidate leukemic stem cells in blast-crisis CML. *N. Engl. J. Med.*, **351**, 657–667.

Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

Lee,K.E. *et al.* (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

Linghu,B. *et al.* (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.*, **10**, R91.

Madigan,D. and Raftery,A. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Stat. Assoc.*, **89**, 1335–1346.

McWeeney,S.K. *et al.* (2010) A gene expression signature of CD34+ cells to predict major cytogenetic response in chronic-phase chronic myeloid leukemia patients treated with imatinib. *Blood*, **115**, 315–325.

Melo,J.V. and Barnes,D.J. (2007) Chronic myeloid leukaemia as a model of disease evolution in human cancer. *Nat. Rev. Cancer*, **7**, 441–453.

Michiels,S. *et al.* (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Mitchell,T. (1997) *Machine Learning*. McGraw Hill, New York.

Mohty,M. *et al.* (2007) The polycomb group BMI1 gene is a molecular marker for predicting prognosis of chronic myeloid leukemia. *Blood*, **110**, 380–383.

Mohty,M. *et al.* (2008) Association between BMI-1 expression, acute graft-versus-host disease, and outcome following allogeneic stem cell transplantation from HLA-identical siblings in chronic myeloid leukemia. *Blood*, **112**, 2163–2166.

Neviani,P. *et al.* (2005) The tumor suppressor PP2A is functionally inactivated in blast crisis CML through the inhibitory activity of the BCR/ABL-regulated SET protein. *Cancer Cell*, **8**, 355–368.

Nguyen,D.V. and Rocke,D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Oehler,V.G. *et al.* (2009a) The preferentially expressed antigen in melanoma (PRAME) inhibits myeloid differentiation in normal hematopoietic and leukemic progenitor cells. *Blood*, **114**, 3299–3308.

Oehler,V.G. *et al.* (2009b) The derivation of diagnostic markers of chronic myeloid leukemia progression from microarray data. *Blood*, **114**, 3292–3298.

Passweg,J.R. *et al.* (2004) Validation and extension of the EBMT Risk Score for patients with chronic myeloid leukaemia (CML) receiving allogeneic haematopoietic stem cell transplants. *Br. J. Haematol.*, **125**, 613–620.

Pavlu,J. *et al.* (2010) Optimizing patient selection for myeloablative allogeneic hematopoietic cell transplantation in chronic myeloid leukemia in chronic phase. *Blood*, **115**, 4018–4020.

Quintas-Cardama,A. and Cortes,J. (2009) Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood*, **113**, 1619–1630.

Radich,J.P. *et al.* (1995) Polymerase chain reaction detection of the BCR-ABL fusion transcript after allogeneic marrow transplantation for chronic myeloid leukemia: results and implications in 346 patients. *Blood*, **85**, 2632–2638.

Radich,J.P. *et al.* (2001) The significance of bcr-abl molecular detection in chronic myeloid leukemia patients "late," 18 months or more after transplantation. *Blood*, **98**, 1701–1707.

Radich,J.P. *et al.* (2003) HLA-matched related hematopoietic cell transplantation for chronic-phase CML using a targeted busulfan and cyclophosphamide preparative regimen. *Blood*, **102**, 31–35.

Radich,J.P. *et al.* (2006) Gene expression changes associated with progression and response in chronic myeloid leukemia. *Proc. Natl Acad. Sci. USA*, **103**, 2794–2799.

Raftery,A.E. (1995) Bayesian model selection in social research (with discussion). *Sociol. Method.*, **25**, 111–196.

Ramaswamy,S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Ren,R. (2005) Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat. Rev. Cancer*, **5**, 172–183.

Sawyers,C.L. *et al.* (2002) Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. *Blood*, **99**, 3530–3539.

Shah,N.P. (2008) Advanced CML: therapeutic options for patients in accelerated and blast phases. *J. Natl Compr. Canc. Netw.*, **6** (Suppl. 2), S31–S36.

Sokal,J.E. *et al.* (1984) Prognostic discrimination in "good-risk" chronic granulocytic leukemia. *Blood*, **63**, 789–799.

Stirewalt,D.L. *et al.* (2008) Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer*, **47**, 8–20.

Tibshirani,R. (1996) Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. B*, **58**, 267–288.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

van 't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Xu,J.Z. and Wong,C.W. (2010) Hunting for robust gene signature from cancer profiling data: sources of variability, different interpretations, and recent methodological developments. *Cancer Lett.*, **296**, 9–16.

Yang,A.J. and Song,X.Y. (2010) Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, **26**, 215–222.

Yeung,K.Y. *et al.* (2005) Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, **21**, 2394–2402.

Yong,A.S. *et al.* (2006) Molecular profiling of CD34+ cells identifies low expression of CD7, along with high expression of proteinase 3 or elastase, as predictors of longer survival in patients with CML. *Blood*, **107**, 205–212.

Zhao,C. *et al.* (2009) Hedgehog signalling is essential for maintenance of cancer stem cells in myeloid leukaemia. *Nature*, **458**, 776–779.