1-1-2013

# Identification of SNP-containing regulatory motifs in the myelodysplastic syndromes model using SNP arrays ad gene expression arrays

Jing Fan

Jennifer G. Dy

Chung-Che Chang
*University of Central Florida*

Xiaoboo Zhou

## Recommended Citation

Fan, Jing; Dy, Jennifer G.; Chang, Chung-Che; and Zhou, Xiaoboo, "Identification of SNP-containing regulatory motifs in the myelodysplastic syndromes model using SNP arrays ad gene expression arrays" (2013). *Faculty Bibliography 2010s*. 3960.
https://stars.library.ucf.edu/facultybib2010/3960

Original Article

# Identification of SNP-containing regulatory motifs in the myelodysplastic syndromes model using SNP arrays and gene expression arrays

Jing Fan[1], Jennifer G. Dy[1], Chung-Che Chang[2] and Xiaobo Zhou[3]

**Abstract**

Myelodysplastic syndromes have increased in frequency and incidence in the American population, but patient prognosis has not significantly improved over the last decade. Such improvements could be realized if biomarkers for accurate diagnosis and prognostic stratification were successfully identified. In this study, we propose a method that associates two state-of-the-art array technologies—single nucleotide polymorphism (SNP) array and gene expression array—with gene motifs considered transcription factor–binding sites (TFBS). We are particularly interested in SNP-containing motifs introduced by genetic variation and mutation as TFBS. The potential regulation of SNP-containing motifs affects only when certain mutations occur. These motifs can be identified from a group of co-expressed genes with copy number variation. Then, we used a sliding window to identify motif candidates near SNPs on gene sequences. The candidates were filtered by coarse thresholding and fine statistical testing. Using the regression-based LARS-EN algorithm and a level-wise sequence combination procedure, we identified 28 SNP-containing motifs as candidate TFBS. We confirmed 21 of the 28 motifs with ChIP-chip fragments in the TRANSFAC database. Another six motifs were validated by TRANSFAC via searching binding fragments on co-regulated genes. The identified motifs and their location genes can be considered potential biomarkers for myelodysplastic syndromes. Thus, our proposed method, a novel strategy for associating two data categories, is capable of integrating information from different sources to identify reliable candidate regulatory SNP-containing motifs introduced by genetic variation and mutation.

**Key words** Association study, genetic variation and mutation, transcription factor–binding sites, myelodysplastic syndromes

Myelodysplastic syndromes (MDS) are a heterogeneous group of clonal hematopoietic disorders characterized by peripheral cytopenia, morphologic dysplasia, and susceptibility to leukemic transformation[1,2]. Copy number variation detection, genotyping, and related studies of MDS with single nucleotide polymorphism (SNP) array are powerful tools for molecular karyotyping that have increasingly been used in recent years [1,2]. Gene expression array is also employed to identify genetic biomarkers for MDS[3,4]. However, MDS arises from diseased stem cells (blasts) that induce dysplastic hematopoiesis of multiple cell lineages, including erythroid lineage, which leads to anemia; myeloid lineage, which leads to neutropenia/myeloid leukemia transformation; and megakaryocytic lineage, which leads to thrombocytopenia. That multi-lineage involvement has seriously hindered success in previous studies using SNP or expression microarray alone. Because genes showing both allelic imbalance and differential expression profiles—as determined by SNP array and expression

**Authors' Affiliations:** [1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA; [2]Department of Pathology, Florida Hospital, University of Central Florida, Orlando, Fl 32803, USA; [3]Department of Radiology, The Methodist Hospital Research Institute, Weill Medical College & Cornell University, Houston, TX 77030, USA.

**Corresponding Author:** Jing Fan, Department of Electrical and Computer Engineering, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA. Tel: +1-857-891-1728; Fax: +1-617-373-8970; Email: fan.j@husky.neu.edu.

**doi:** 10.5732/cjc.012.10113

array—could be valid targets for building biomarkers, we have integrated these two types of arrays for mechanistic and bioinformatic studies of MDS. One possible way that potential biomarkers work is through the regulatory pathway using SNP-containing transcription factor-binding sites (TFBS) and other variation-related mechanisms. These two state-of-the-art technologies, high-throughput genome-wide profiling using SNP array and gene expression (oligonucleotide) microarrays, have been associated and studied in other disease models[5-9] but not yet, to our knowledge, in the MDS model. In this study, we demonstrated an association analysis method that integrates information from two types of arrays and use this approach to identify genetic variation-induced regulation events. This approach considers SNP-containing motifs potential biomarkers that change the normal/standard regulatory events by functioning as TFBS. It takes advantages of both gene expression array and SNP array to identify differentially expressed genes as well as genes containing SNP sites to find potential biomakers in a data association study manner.

Current association studies can be categorized into three classes: 1) gene-based association studies[10,11], 2) region-based association studies[8,9,12], and 3) whole genome association studies[13-15]. Gene-based association studies are the most widely used approaches and are effective in identifying genes that cause disease when the underlying genetic defect, considered a genetic marker, resides in a specific biological pathway. On the other hand, region-based association studies are successful when a connection between disease phenotype and specific regions on a chromosome is established. The region-based approach can be used to identify novel genes by studying genetic markers in a target region with a certain density of SNPs to locate a disease-associated locus. Whole genome association studies extend the scope of region-based research for discovery of novel genes regulated genome-wide. Whole genome studies do not require a prior knowledge of candidate genes. A map between gene expression or phenotype traits and genetic locus can be established genome-wide. The goal of all three classes is to identify genetic markers that are possible disease-casual regions or specific alleles to perform popular linkage studies.

In this paper, we propose a method that associates gene profiles and genetic variation or mutation in a motif-based manner. Our method can be categorized as a genome-wide association study. However, we focused on specific chromosomes—chromosomes 5, 7, and 8—that are considered highly suspicious disease causal chromosomes and show abnormal behaviors in MDS patients[2,16]. We employed the connection between gene expression, provided by gene expression array, and TFBS that contain genetic variations, identified by SNP array, to associate two types of arrays. Transcription is well known to be regulated through TFBS or motifs[17]. In the current literature, the most popular computational methods for TFBS identification are regression-based methods[18-20]. These methods represent gene expression as a linear combination[20] or logic regression[19] coefficients of candidate motifs, and the best subset of candidates are selected with feature selection methods or sparsity penalization during regression. Based on the knowledge that transcriptional regulation in eukaryotic organisms requires cooperation of multiple transcriptional factors, the ideal method for selecting potential TFBS or regulatory motifs should be capable of identifying groups of cooperating candidates. Keles *et al.* [19] applied logic regression to address this problem, but the binary restriction of logic regression limited method power by forcing motif frequency to be binary indicators of occurrence, resulting in information loss. The strategy of linear regression followed by variable selection suffered high computational complexity, and different feature selection methods yielded distinct results, which made the problem more confusing.

Our proposed strategy integrates SNP array information into gene and candidate motif selection while addressing limitations of these approaches. Increasing numbers of studies suggest that genetic variations at SNPs and mutations are driving factors for various diseases[21-23]. To associate gene expression profile and SNP information, we selected differentially expressed genes with covariant copy number gain or loss. Then, unlike most existing methods on regulatory TFBS identification, we did not perform exhaustive motif search but instead focused on SNP-containing motifs introduced by genetic variation and mutation identified by SNP array. These motifs are in the neighborhoods of SNPs and may not exist unless the variations or polymorphisms occur. In other words, instead of the general regulatory mechanism induced by transcription factors, we were particularly interested in regulatory events introduced by SNP genetic variations. Therefore, the SNP-containing motifs that act as TFBS reveal novel regulatory events that cannot be identified using traditional motif search and selection methods. Followed by SNP-containing motif candidate search, we employed least angle regression via elastic net (LARS-EN)[24] to perform regression and derive a sparse solution simultaneously. Unlike least absolute shrinkage and selection operator (LASSO)[25], LARS-EN takes the group effect into consideration and selects all good feature components in a group instead of the best one. Thus, it takes the cooperation of selected motifs into consideration and is effective under the circumstance where features are not binary as in logic regression. As mentioned above, model fitting and selection are carried out by LARS-EN at the same time with much smaller computational cost. Our strategy does not conflict with the popular TFBS selection methods. Furthermore, it can be considered an important addition to conventional candidate sets selected by other methods.

The major contribution of this work is that we model the relationship between SNP-introduced TFBS and gene expression levels in the MDS model and identify potential biomarkers as TFBS in regulatory events. As mentioned above, SNP-containing motifs, which potentially regulate genes with genetic variations and mutations, link information from SNP array and gene expression array. This is another contribution of this work to both regulatory motif or binding site studies and association studies. Genotypic variations at SNP positions bring new possible binding sites, which may introduce novel modes of gene regulation, altering gene expression and thereby impacting disease progression. On the other hand, compared with the traditional association study, we did not employ a single SNP or a group of SNPs as a marker; instead, motifs were the surrogate objects to study regulatory events and served as genetic markers as well. It is pellucid that when genetic variations take place at TFBS, gene expression is regulated to change their regular behaviors from its standard behavior. Therefore, in this study, SNP-containing motifs bridged two types of array data together, allowing us to infer a new set of candidate TFBS motifs and from which new regulatory relationship could be introduced.

Instead of experimental verification, we used transcription factor library TRANSFAC Professional (BIOBASE GmbH, Wolfenbuttel, Germany)[26] containing experimentally verified binding site information to validate the effectiveness of our proposed method. We also incorporated gene ontology (GO) information to infer the nature of co-regulated genes to validate newly identified motifs. Three level comparisons—biological process, cellular component, and molecular function—were respectively used to study the similarities of co-regulated genes and to validate our motifs together with TRANSFAC binding site information.

## Materials and Methods

### Experiment materials

Fourteen SNP arrays and 14 gene expression arrays were generated from 14 samples of myeloid, lymphoid, and blast tissue, of which half were disease samples and the other half were controls. Genomic RNA was extracted from each sample with the Qiagen Allprep RNA/DNA Mini Kit (Qiagen Valencia, CA) and stored at −80℃. The quantity of RNA was measured using NanoDrop ND-1000 (NanoDrop Technologies, Wilmington, DE), and the quality of RNA was further assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Foster City, CA) according to the manufacturer's instructions. The flow-through was used for total RNA extraction with the RNeasy Micro Kit (QIAGEN) according to the manufacturer's protocol to ensure the recovery of adequately concentrated RNA for expression array. Genotyping was performed using 250K *Nsp*I SNP-microarray chips (Affymetrix, UK) and processed according to the manufacturer's instructions. Samples of 250 ng genomic DNA were digested with *Nsp*I for 2 h at 37℃, followed by adaptor ligation, polymerase chain reaction (PCR) amplification, fragmentation, labeling, and hybridization. Then, 3 μL of PCR product and 4.5 μL of fragmentation product were eletrophoresed to confirm DNA processing. The Affymetrix 2-cycle amplification protocol was used for expression array analysis using 10 ng RNA to start. The Affymetrix 450 fluidics station and the Affymetrix gene scanner were used to wash, stain, and scan the arrays.

### Gene array processing and gene selection

The raw data of gene expression were summarized using the robust multi-array analysis (RMA) algorithm[27], a component of the Affymetrix Expression Console software (Affymetrix, CA, US). RMA starts with a non-linear background correction based on a single chip to detect the perfect match signal. Then, multiple chips are normalized and analyzed jointly to make the distribution identical across arrays. Following the normalization step, summarization translates perfect match values into expression measures. RMA assumes the probe affinity effects sum to zero and estimates the gene effect using median polishing to avoid outlier probes. This algorithm is computationally efficient and consistently variant in individual signal. We processed the MDS gene microarray data with RMA to take the advantage of cross-array normalization and low computation intensity in log scale.

Copy number variation and genotyping call are two categories of information provided by SNP arrays. Affymetrix Genotyping Console (Affymetrix, CA, US) was used to extract copy number variation and genotyping call from raw intensity files. The algorithm to summarize copy number is based on hidden Markov model, which identifies the dynamics of genomic copy number changes. The Viterbi-algorithm-based method is resilient to local copy number perturbations and is able to find globally optimal results. The non-diploid state deletion and amplification indicate either disease susceptibility or resistance; thus, copy number variation explains not only genetic variation but also certain types of diseases. On the other hand, Bayesian robust linear model with Mahalanobis distance classifier (BRLMM) was employed in the Genotyping Console software to infer the geno-typing. In our study, accurate identification of genotyping for each SNP was of significance in that this information would be utilized to identify SNP-containing motifs, a major component of this work. Similar to the RMA algorithm, BRLMM is also a multi-chip-based method that allows simultaneous estimation of both probe and allele signals for each SNP. In addition, by introducing multiple sample classification, the genotyping estimation takes advantage of several SNPs other than the current one to be more robust. Bayesian prior information further enables the desired performance of BRLMM by integrating posterior estimation. Genetic variant information provided by copy number variation and genotyping call is then linked to phenotypes and different traits of samples in later sections.

After translating probe intensities into genetic traits of gene expression and SNPs, statistical analysis was then performed on expression data using significance analysis of microarrays (SAM) to select differentially expressed genes. SAM selects genes with statistical significance according to a score describing the relationship between gene expression and standard deviation of repeated measurements[28]. Both up-regulated and down-regulated genes with equal or larger magnitude of scores were identified for further study; these genes that were differentially regulated between the control and disease groups were considered important in the mechanism of disease.

SNPs are associated with genes if they are located in the expanded gene region—the region containing the gene as well as 20 kb upstream and 1 kb downstream. The upstream region is large enough to include promoters, which contain TFBS and therefore regulate gene transcription. The downstream extension includes untranslated regions, which are dominant contents on gene expression arrays. We denoted SNPs located in the expanded gene region as SNPs related to the gene.

Because copy number variation regions are target regions of MDS research, the gene list containing up-regulated and down-regulated genes could be further shortened by integrating copy number variation information. For each SNP located within an amplified expanded gene region, we determined whether the expression files were associated with copy number variations. It is equivalent to the case in which copy number status was determined to be in either the diploid or non-diploid state for each differentially expressed gene. If copy number was covariant with expression, the gene was reserved for further application, whether there was a gain or loss in copy number. These reserved genes with "oncogene-like" properties indicated dysregulated mechanisms other than gene dosage alteration. Therefore, the genetic trait copy number decreased the size of candidate gene set by filtering genes with expression files uncorrelated to genetic copy number variations.

Following the previous stringent filtering, a more parsimonious set of genes was regarded as the potential candidate set for MDS. These genes were clustered according to their expression profile; thus, genes that behaved similarly across the sample set were clustered as a group of co-expressed genes. The hierarchical clustering algorithm[29] represents the data in a tree structure without prior information of number of clusters. The cut-off value is a threshold of similarity measurement between each pair of branches across the dendrogram, which thus separates the dendrogram into several clusters with a similarity value higher than the cut-off. For each cluster with a group of co-expressed genes with similar behavior, we then developed a method to find the regulatory mechanism with respect to SNPs.

## SNP-containing motif search

Genotyping call represents true sequence composition at SNP positions. The genotype is the specific allele constitution of an individual SNP that comprises most genetic variations. Variant genotypes are the dominant reason for variety of phenotypes. SNPs with different allele compositions in control and disease samples are highly suspected to carry potential factors that are closely associated with disease pathogenesis and could be considered biomarkers. Typically, there are four different types of genotyping call from SNP arrays: AA, BB, AB, and no call. For each gene, we had its flank expanded gene region and genotypes of its related SNPs, allowing us to obtain accurate individual genotypes.

We integrated allelle genotype information into the expanded gene region by introducing the terminology "twin expanded gene region." Twin expanded gene region refers to two expanded gene regions whose allele constitutions are different only at AB call SNPs. The standard sequences of genes and their flanks are available in all popular gene databases from which expanded gene regions are obtained. For each sample, we first duplicated each expanded gene region in forms of standard sequence; that is, two twin sequences were created. For AA call and BB call SNPs, the identified genotyping calls replace allele contents in both twin sequences at SNP positions. On the other hand, for AB calls, allele contents at SNP positions of twin sequences are substituted by A and B, respectively. In the no call case, which is rare, SNP compositions retain their contents as obtained from database. With this procedure, genotyping information was integrated into twin sequences for each gene and expanded gene region. Therefore, a pair of twin expanded gene regions of a single gene was identical except for the AB call SNP alleles. We illustrated the concept of expanded gene region in Figure 1. In this example, if the allele on gene A appears in form of AA call or BB call as G, the motif candidate might not be considered because of the low frequency. In other words, the AB call enables some of the short sequence to be a candidate TFBS.

We collected twin sequences of different samples of each gene in a co-expressed group into a single sequence set. More specifically, assume we have $n$ instances and $k$ genes in a co-expressed cluster. Then the size of combined sequence set is $2nk$. Motif search was employed to find shared patterns across the set. Here, we were interested in the SNP-containing motifs that include SNPs in the shared patterns. It was convenient to separate the sequence set into two subsets: disease and control. In addition, because we assumed the variations and mutations cause MDS, we considered the SNPs in the disease group only.

Assuming the motif length is $L$, then for each SNP, $L$-$1$ upstream flank, $L$-$1$ downstream flank, and SNP position were extracted to be a short sequence of length $2L$-$1$. A window of length $L$ slides over the short sequence, and sequences within the window are stored. Thus for each SNP, we obtained $L$-$1$ sequences with length $L$ as candidate motifs. This procedure is illustrated by Figure 2. Then we repeated the same procedure for all SNPs in the combined sequence subset of disease samples. These $L$-length sequences carried at least one SNP and were considered the candidate SNP-containing motifs. We then searched frequencies of candidate motifs in the combined sequence set of disease and control obtained earlier. Because there were pairs of twin sequences and AA call, BB call and no call, there were some identical candidate motifs. We kept the unique candidates and calculated their frequencies. Assuming there are $M_0$ candidate motifs and if we consider frequency as features, the size of the feature matrix is $2nk \times M_0$.

We had a pair of twin expanded gene regions for each gene of a certain sample. For the related SNPs of each gene, frequencies of elements in a twin pair are very similar and sometimes identical. We took average frequencies of SNPs for a pair of twin sequences, and thus the size of feature matrix was $nk \times M_0$. This step eliminated the duplicate effect of twin sequences. Because of the existence of AB call, it is possible that certain frequencies are decimals instead of integers.

The desired candidate motifs should spread out among different genes. To identify TFBS from these candidates, we first filtered candidate motifs with frequencies smaller than a given threshold, which was equal to the number of samples. This procedure is a coarse process that removes candidates highly unlikely to be TFBS. Then, a fine statistical process was performed based on the filtered frequency matrix with statistical testing. Here, we employed the similar method of hypothesis testing in microarray experiments[30]. The random perturbation generates sequences with the same length and ATGC proportions as original genes. Assuming that the number of permutation is $N$ and that there are $n$ genes in a co-expressed set, we randomized the sequence order and rearranged the sequence according to the random order for each permutation. Then we repeated the same procedure for each permutation as previously described to generate the averaged frequency matrix. In comparing the frequencies of actual gene sequences and randomly permutated sequences, the null hypothesis is that the summation of averaged frequencies of a motif in the real gene sample set is consistent with the motif distribution of the summation of averaged frequency of randomized sequences. The test statistics are the summation of

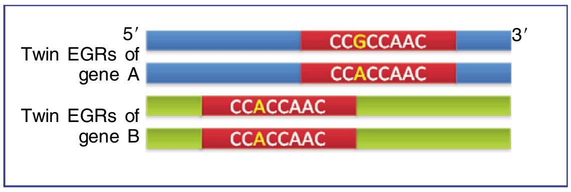

Figure 1. **Illustration of twin expanded gene regions (EGRs).** Here, single nucleotide polymorphisms (SNPs) introduce potential novel motifs. For one gene, the potential motif (red box) has exactly the same DNA sequence except the heterozygous SNP position (G and A) in gene A. The motif will be missed if the heterozygous SNP position is G.

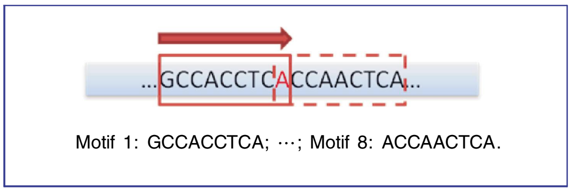

Figure 2. **Example of how sliding window selects candidate motifs with length $L$= 9 from a sequence fragment.** The SNP is noted in red. A total of 8 candidate motifs are generated while the sliding window cruises from left to right.

averaged frequencies for different motifs. For two-side alternative hypotheses, we calculated the *P* value as follows:

$$p_j = \frac{\sum_{i=1}^{N} I(|t_{j,i}| \geqslant |t_j|)}{N} \qquad (1)$$

For the $i$-th permutation, $i = 1, \cdots, N$, we computed the test statistics $t_{i,1} \cdots t_{i,m}$ for each null hypothesis $H_j$, $j = 1, \cdots, m$, where $m$ is the number of motifs after coarse filtering. $I$ is the indicator function, which is equal to 1 if the condition is satisfied and 0 otherwise.

Given that the null hypothesis is true, the *P* value is the probability of getting a statistical testing value as extreme as the observed instance. We selected motifs with *P* value smaller than 0.05 to indicate statistical significance. This fine filtering step reduces the size of candidate motifs by removing motifs occurring as random incidents. The whole procedure of motif search and candidate motif frequency matrix calculation is illustrated in Figure 3.

In this study, we constrained the length of SNP-containing motifs from 6 to 11 bp, but the range could be easily extended. However, a longer length is accompanied with more special sequences, which are sparse in a co-expressed gene group. In our study, with a group of 10 genes, the length of candidate motifs ranged from 6 to 10, and motifs with length of 11 bp were filtered mainly in the coarse thresholding step due to the sparsity of the frequency matrix. Increasing the length of motifs would require enlarging the number of co-expressed genes in a cluster. This could only be achieved by relaxing the rule of differentially expressed gene selection or loosening the similarity measurement in the gene cluster group, both of which would falsely detect genes and result in false positives during SNP-containing motif identification.

## Association model with regression

With gene expression data and averaged motif frequency matrix of corresponding genes, it is natural to fit a regression model to discover the latent relationship between motifs as variables and expression as response. Regression is capable of causal relationship modeling, prediction, and inference without knowledge of the underlying procedures that produced the data. In our studies, we aimed to model the causal relationship and more specifically, the regulatory mechanism between gene expression and SNP-containing motifs generated from SNP array. Thus, the regression model can be presented as follows:

$$\begin{gathered} y_i = \mathrm{a}_1 X_{i1} + \mathrm{a}_2 X_{i2} + \cdots + \mathrm{a}_j X_{ij} + b_i \\ i = 1, 2, \cdots, K; j = 1, 2, \cdots, M \end{gathered} \qquad (2)$$

For simplicity, we present this formula with only one example, but the model is easily extended to multiple samples. In the simplified formula, $y_i$ is the expression value of $i$-th gene in the co-expressed gene cluster; $X_{ij}$ is appearance frequency of $j$-th SNP-containing motif in $i$-th gene; $\mathrm{a}_j$ is the weight or regression coefficient of $j$-th motif; and $b_i$ is incorporated to perform as other possible regulatory mechanisms or other TFBS effects not

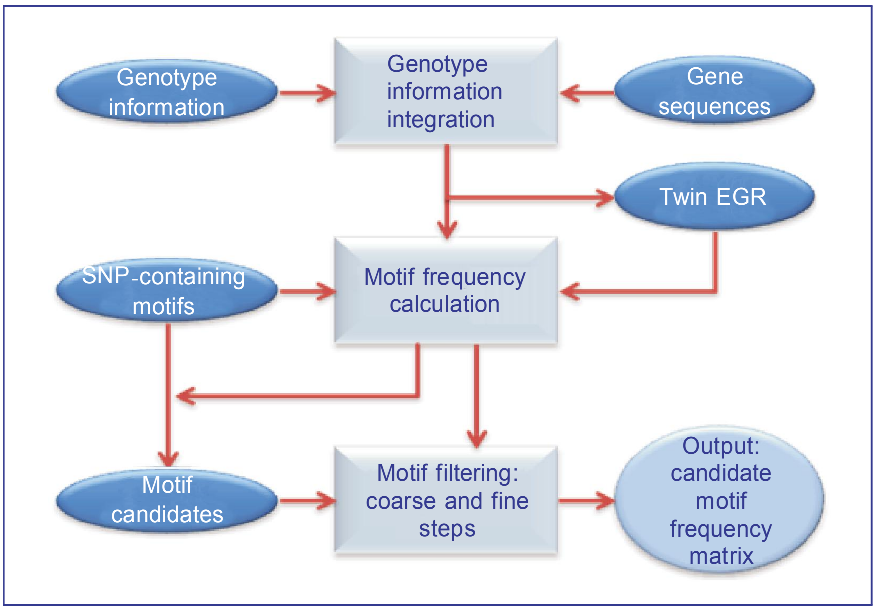
Genotype information
Genotype information integration
Gene sequences
Twin EGR
SNP-containing motifs
Motif frequency calculation
Motif candidates
Motif filtering: coarse and fine steps
Output: candidate motif frequency matrix

Figure 3. **Flowchart of the SNP-containing motif frequency matrix calculation.** Details are provided in the Method section.

included in candidate set.

As discussed above, SNP-containing motifs are short sequences with a certain length, which, in our studies, ranged from 6 to 11 bp and contained SNPs at least at one of its location genes. In other words, it is unnecessary or impossible that all the motif location genes contain the same SNP. For example, for a certain motif, only one of its location genes contains the SNP whereas other locations are sequences without SNPs on shared pattern. Another case is that the SNP enables the motif to appear in the gene that may not contain the motif unless specific allele expression occurs at the SNP position. This indicates that the gene variations at SNPs may introduce a motif that is not originally contained in the gene as a regulatory factor. This is illustrated by Figure 2. Another aspect of the mutation events is that the gene variation at SNPs increases the frequency of a certain motif and leads to changes in gene expression. Under this circumstance, if the variation in frequency across the co-expressed gene cluster is consistent with the variation in gene expression, this motif is highly probable to be a regulatory factor. Therefore, it is unnecessary for the same SNP-containing motif to have that exact SNP at all its located genes across the co-expressed gene cluster.

## Motif selection and ranking

With the previously described regression model, any fitting process that generates coefficients can be adopted, such as ordinary least square error base method, ridge regression with a $L_2$ norm penalization and LASSO with a $L_1$ norm term[25]. In addition to residual sum of square errors, prediction accuracy and interpretation capability of the model are also key elements to evaluate a fitting procedure. Unfortunately, ordinary least square method suffers both prediction accuracy and interpretation ability problems. Ridge regression fails to achieve a parsimonious set of predictors, which fails to provide good interpretation of model, though it reaches higher prediction accuracy than ordinary least square regression. LASSO, in which $L_1$ norm penalization term helps correct both issues, however, suffers other problems: 1) it fails in the $p>n$ case in which the number of features $p$ is larger than that of samples $n$, and 2) it does not consider the grouping effect and is vulnerable to select only a single feature from a highly correlated desired feature group.

Instead of these procedures, a surrogate strategy is LARS-EN (least angle regression via elastic net [24]), which performs continuous shrinkage and automatic variable selection simultaneously. LARS-EN is a compromise between ridge regression and LASSO. It includes $L_2$ norm penalization from ridge regression to shrink coefficients of correlated features towards each other by allowing them to borrow power from each other[24], and it also includes $L_1$ norm penalization from LASSO to guarantee the sparsity. It handles the $p > n$ case well. In our case, we had a much larger number of SNP-containing motif candidates than co-expressed genes of different samples. In addition, the grouping effect enables the algorithm to select all desired variables from a highly correlated feature group. Moreover, the magnitudes of regression coefficients are employed as the criteria to rank the SNPs in both models.

The elastic net coefficient is a proven scaled version of the naïve elastic net [24]. The naïve elastic net estimator is defined as:

$$\hat{a} = \arg\min_{a}\{L(\lambda_1, \lambda_2, a)\} \tag{3}$$

where

$$L(\lambda_1, \lambda_2, a) = |y - Xa|^2 + \lambda_1 |a|_1 + \lambda_2 |a|^2 \tag{4}$$

$$|a|^2 = \sum\nolimits_{j=1}^{M} a_j^2 \quad |a|_1 = \sum\nolimits_{j=1}^{M} |a_j|$$

$\lambda_1$ and $\lambda_2$ are fixed positive weights that put different emphasis on $L_1$ norm and $L_2$ norm penalization to balance the sparsity and the grouping effect. Thus, the method possesses the desired properties of both LASSO and ridge regression.

Solving â in equation 3 is equivalent to optimize the following problem:

$$\hat{a} = \arg\min_{a}\{|y - Xa|^2\}, \quad \text{subject to } (1-\gamma) \\ |a|_1 + \gamma |a|^2 \leqslant t \text{ for some } t \tag{5}$$

Where $\gamma = \lambda_2 / (\lambda_1 + \lambda_2)$ and $(1-\gamma)|a|_1 + \gamma |a|^2$ is the elastic net penalty. Least angle regression (LARS) was employed to resolve elastic net solution path to efficiently find the estimators. The relationship between elastic net estimator and naïve elastic net estimator is

$$\hat{a}(\text{elastic net}) = (1+\lambda_2)\hat{a}(\text{naive elastic net}) \tag{6}$$

To make the computation feasible and efficient, the response $y$, which is a vector of gene expressions in our case, is centered. The features are standardized as follows:

$$\sum_{i=1}^{K}\sum_{p=1}^{N} y_i^p = 0 \sum_{i=1}^{K}\sum_{p=1}^{N} X_{ij}^p = 0 \sum_{i=1}^{K}\sum_{p=1}^{N} (X_{ij}^p)^2 = 1$$

$$\text{for } j = 1, 2, \cdots, M \tag{7}$$

Let $y = (y^1, y^2, \cdots, y^N)^T$ be the gene expression profile of every gene in the co-expressed clusters of N samples, where $y^p = (y_1^p, y_2^p, \cdots y_K^p)^T$ , $p$ = 1, 2, ⋯, $N$

are gene expressions of sample $p$. $\mathbf{X}=(\mathbf{X}_1\,|\,\mathbf{X}_2\,|\cdots|\,\mathbf{X}_M)$ is the feature matrixthat provides frequency information for each motif, where $\mathbf{X}_j=(X_{1j}^1,\cdots X_{Kj}^1,X_{1j}^2,\cdots X_{Kj}^2,\cdots,X_{1j}^N,\cdots X_{Kj}^N)^T$ is the feature vector that describes the frequency of a certai n motif of co-expressed genes across all samples. $X_{ij}^p$ is the frequency of motif $j$ appearing in the gene $i$ in the $p$-th sample.

The LARS-EN algorithm finds solutions with any user-defined sparsity, which is equal to the desired size of the selected features. In our study, prior information on the size of the selected motif set was unknown. Therefore, cross validation was employed to estimate the optimal size of selected motifs. For each fixed $\lambda_2$, the tuning parameter s is chosen by 10-fold cross validation to derive the smallest cross validation error. $s$ is standardized bound represented as $s=t\,/\sum_{j=1}^{M}|\hat{a}_j^o\,|$, which is an indication of the fraction of the $L_1$ norm. $\hat{a}_j^o$ is the ordinary least square regression coefficient. The size of optimal subset of motifs is calculated by multiplying *s* with the number of total steps of optimization.

The LARS-EN method identifies â sparse solution with good prediction accuracy, encourages the grouping effect, and introduces a clear interpretation of the model. Non-zero elements in â indicate the effect of features, which is the regulatory effect in our case, and the magnitude denotes the significance or how strong the motif regulates the gene expression.

## Motif combination

Because a sliding window was used to build the candidate set of motifs of different lengths, overlaps may exist between motifs selected by LARS-EN. Similar to secondary structure prediction of protein sequences[31], a level-wise motif combination strategy was employed here to remove highly overlapped motifs. In the LARS-EN selected motif candidate set, if a motif $M$=$c_1$, $c_2$, $\cdots$, $c_n$ and both of its sub-patterns $m_1$=$c_1$, $\cdots$, $c_{n\text{I}1}$ and $m_2$=$c_2$, $\cdots$, $c_n$ are included in the same set, we discarded the two shorter sub-patterns and kept the longer motif. The sub-patterns $m_i$,$i$=$1$, $2$ and motif $M$ were kept if only one subsequence appeared at any end of the motifs. Then, motif length grew further until it reached the upper length limit.

# Results

## Selected gene cluster

Recent studies reported that abnormal behaviors (for example, copy number gain or loss) on sections of chromosomes 5, 7, and 8 are highly probable as disease causal factors for MDS. We focused on chromosome 7 and selected one co-expressed gene cluster. There were 54,675 probes representing genes with duplications on gene expression array across the genome. We selected 397 probes that satisfied both differential expression, with *P* value less than 0.05, and covariance, with copy number variation on chromosome 7. Then, 98 probes were selected in one cluster according to their expression similarities, with a cut-off value at 0.83 for the hierarchical cluster. To reveal the regulatory effect of SNPs, prerequisites of the number of related SNPs located in a gene's expanded gene region were used to filter some probes, and the threshold of SNP quantity we used here was 6. Then, 10 genes represented by 22 probes on the gene expression array were identified, and the expression values of duplicated probes of the same gene were averaged. The brief functional description, length, and number of related SNP for each gene are listed in Table 1.

Figure 4 shows every selected gene, their expanded gene region and related SNPs, and their relative positions, as well as box plots of copy number around each SNP in 7 disease samples. Copy number gain was only observed at gene *MDH2*, whereas other genes consistently suffered copy number loss in disease samples. The copy number change events on these differentially expressed genes indicate there are some biological insights, such as different regulatory effects from control cases, which can help biologists further explain the mechanism of MDS.

## Selected SNP-containing motifs

Sliding window generated 8,340 unique SNP-containing motif candidates on twin expanded gene regions of different samples with length ranging from 6 to 11 bp. Then candidates that did not widely spread were discarded. The motifs that appeared more than twice in each sample were reserved for further use. After the coarse thresholding step, statistical testing with random permutation sequences selected 2,159 candidates with *P* value less than 0.05. At this step, the number of motifs was highly reduced, compared with the original number, which lessened the computational cost.

With a parsimonious set of candidate motifs, we predicted the causal relationship between the motif frequency and gene expression profiles with LARS-EN. The parameter setting in LARS-EN followed the standard strategy in the study by Zou *et al.*[24]. In our case, we set $\lambda_2$ to 1,000 because of the large number of motifs (features) and implemented cross validation to identify the best size of the optimal subset. Finally, 30 motifs were selected and are listed in Table 2. The length of these selected motifs ranges from 6 to 10 bp. Motifs

Table 1. **Ten selected genes and their basic information**

| Gene name | Description | Length | Number of SNPs |
|---|---|---|---|
| *EIF2AK1* | Eukaryotic translation initiation factor 2-alpha kinase 1 | 36,906 | 7 |
| *MCA2_HUMAN* | Multisynthetase complex auxiliary component p38 | 14,583 | 6 |
| *CBX3* | Chromobox protein homolog 3 | 12,195 | 6 |
| *TES* | Testin (TESS) | 48,255 | 12 |
| *CALU* | Calumenin precursor | 32,092 | 7 |
| *HBP1* | High mobility group (HMG) box-containing protein 1 | 33,514 | 7 |
| *MDH2* | Malate dehydrogenase, mitochondrial precursor | 18,536 | 7 |
| *WNT2* | Protein Wnt-2 precursor | 46,062 | 9 |
| *PIK3CG* | Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit gamma isoform | 41,642 | 6 |

SNP, single nucleotide polymorphism.

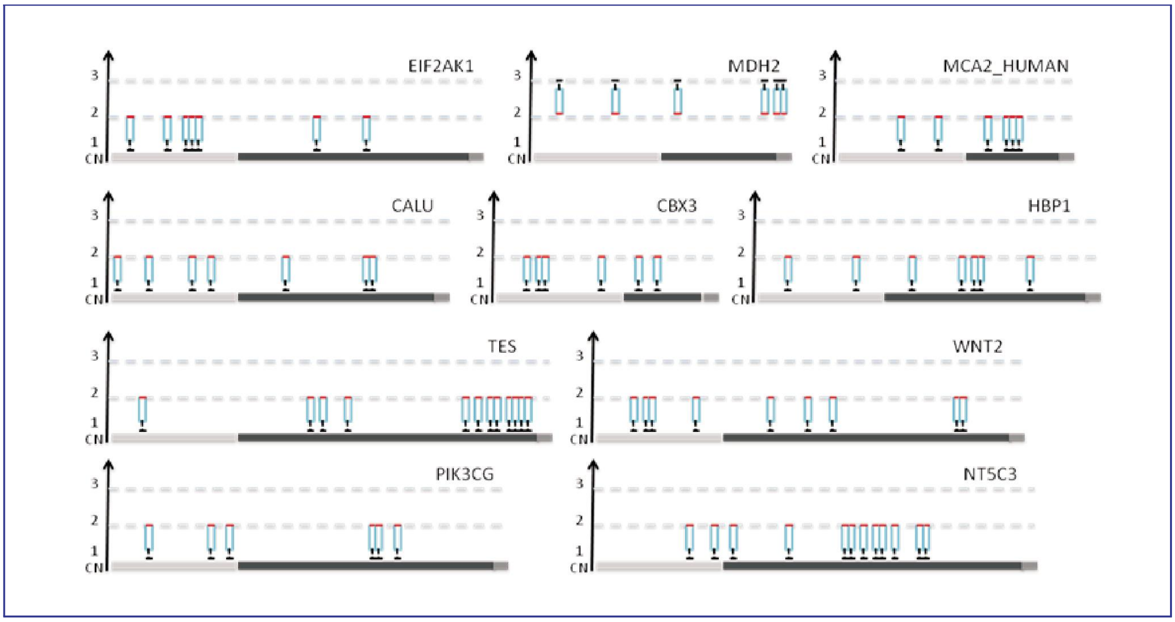

Figure 4. **Visualization of the expanded gene region (EGR) of a gene and its related SNP.** The box plots describe copy numbers of each SNP across 7 disease samples.

TTTCAC and TTCACT were discarded according to level-wise motif combination rules because high-level motif TTTCACT appears in the selected set.

In Table 2, we listed regression coefficients, distributions (the number of different genes where motifs locate) and *P* values for each motif based on the random permutation, as well as the binary vector denoting whether a certain motif could be traced in the TRANSFAC database on its location genes. The rank of the 28 selected motifs according regression coefficients is listed in the last column of Table 2.

We used the motif GTGCCAC, which was distributed in all genes in the selected cluster, as an example (Figure 5). This motif was introduced to the set by SNP rs2345060, whose allele polymorphisms are C and T at the end of the motif GTGCCAC. This motif was found in samples S3, S4, S7, S9, S11, S13, and S14, and 4 of its 6 appearances are in MDS cases. In all other samples, the SNP positions are T and the sequence is regarded as GTGCCAT. Thus, the SNP variation introduces the SNP-containing motif that is present in more than half of the disease samples. Therefore, this type of motif may introduce the regulatory effect through TFBS to disease samples. We drew the box plot in Figure 5 to illustrate how this motif distributes in 10 genes across the 14 sample set. The motif was

Table 2. Selected motifs by the proposed method with regression information

| Index | Motifs[a] | Weight | Distribution | *P* value | TRANSFAC | Rank |
|---|---|---|---|---|---|---|
| 1 | AATGGG | 1.040,5 | 10 | < 0.01 | 1 | 22 |
| 2 | CATTGC | 1.128,8 | 10 | < 0.01 | 1 | 21 |
| 3 | TCAGGG | 2.064,7 | 10 | < 0.01 | 1 | 13 |
| 4 | CACTTT | 2.460,3 | 10 | < 0.01 | 1 | 8 |
| 5 | ATTTCTC | 2.172,1 | 10 | < 0.01 | 1 | 12 |
| 6 | TTTCACT | 2.348,5 | 10 | < 0.01 | 1 | 11 |
| 7 | TTCACTT | 2.475,0 | 10 | < 0.01 | 1 | 7 |
| 8 | GTGCCAC | 1.430,1 | 10 | < 0.01 | 1 | 18 |
| 9 | CTGTGTCA | 2.615,9 | 8 | < 0.01 | 1 | 5 |
| 10 | TTTAGAAA | 0.948,3 | 10 | < 0.02 | 1 | 23 |
| 11 | CTGTCACT | 1.693,5 | 8 | < 0.01 | 1 | 17 |
| 12 | GAGTTCCA | 4.330,4 | 9 | < 0.02 | 1 | 2 |
| 13 | TTTTGGAG | 1.742,8 | 9 | < 0.01 | 1 | 15 |
| 14 | AGGAAAAT | 0.116,7 | 10 | < 0.01 | 1 | 25 |
| 15 | TTACTGAG | 2.934,2 | 9 | < 0.01 | 1 | 4 |
| 16 | GGCAGATT | 0.111,7 | 9 | < 0.01 | 1 | 26 |
| 17 | CTTAAAAT | 4.426,2 | 10 | < 0.01 | 1 | 1 |
| 18 | CATGTGAA | 2.382,9 | 10 | < 0.04 | 1 | 10 |
| 19 | TGTGCCAC | 1.730,1 | 9 | < 0.01 | 0 | 16 |
| 20 | AAAAAGAA | 1.379,6 | 10 | < 0.01 | 1 | 19 |
| 21 | CCCTGCAGA | 2.002,4 | 3 | < 0.03 | 0 | 14 |
| 22 | TTTTGGAGT | 1.273,4 | 4 | < 0.03 | 0 | 20 |
| 23 | CTTGCTGCC | 2.549,5 | 5 | < 0.01 | 1 | 6 |
| 24 | AGGCAGATT | 3.328,2 | 6 | < 0.01 | 1 | 3 |
| 25 | CAGGTTCAC | 0.097,3 | 4 | < 0.02 | 0 | 27 |
| 26 | CTCATTTGAC | 2.440,1 | 3 | < 0.01 | 0 | 9 |
| 27 | ATCTCCTGCC | 0.058,5 | 2 | < 0.02 | 0 | 28 |
| 28 | ATTTCATTTT | 0.716,2 | 6 | < 0.01 | 0 | 24 |

[a]TTTCAC (weight, 0.875,0) and TTCACT (weight, 2.822,1) were filtered according the level-wise combination rule, so no detailed information is provided except for the regression weights. Distribution indicates the number of genes the motif located among the 10 co-expressed gene groups. Rank is defined by the absolute value of weight decreasingly.

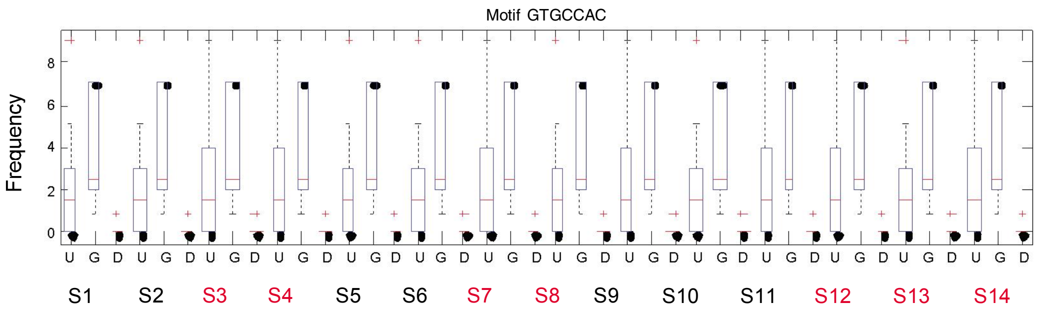

Figure 5. **Box plot of distributions in the upstream gene and downstream regions of motif GTGCCAC across all samples.** The index of MDS samples are marked in red. S stands for samples.

consistently distributed in all samples in the gene region; however, due to the introduction of a SNP, this SNP-containing motif varies in frequency in the upstream region, the most important region for genetic regularity. To conclude, based on the location information of SNP-containing motifs, we observed the situation that these motifs, especially the ones located in the promoter regions, were differentially distributed in control and disease cases and are therefore highly likely to be TFBS or similar regulators of gene expression.

## Validation by TRANSFAC

The TRANSFAC database comprises transcription factors, their experimentally verified binding sites, and the genes they regulate. TRANSFAC professional version 12.1 contains 11,080 factors as well as 141,595 DNA fragments verified by ChIP-chip and related experiments that indicate the *in vivo* binding events of transcription factors. It also provides transcriptional information for 32,296 genes. Among the 10 co-expressed genes in our experiments listed in Table 1, there was no available information for the genes *MCA2*, *EIF2AK1*, or *WNT2* because TRANSFAC failed to provide profiles in *Homo sapiens*. Therefore, the results presented herein describe the other 7 genes.

TRANSFAC did not provide experiment-specified transcription factors for the genes selected in our experiment. Instead, the ChIP-chip validated DNA-binding fragments were used to validate our findings for SNP-containing motifs that appeared in these fragments. We found that 21 of 28 motifs were located in ChIP-chip fragments at least twice. Each fragment is accompanied by a related transcription factor. The number of fragments and suggested transcription factors for each gene are listed in Table 3. All seven selected genes with TRANSFAC records shared transcription factor T00759, and 5 of 7 had T00781 in common. This observation further suggests a relationship between the co-expressed and co-regulated genes.

## Similarity measurement of co-regulated genes

According to TRANSFAC, regulated genes or binding fragments can be assigned to each transcription factor. Of the 4 transcription factors we traced back to the selected 21 motifs from the ChIP-chip fragments, T03828 has 20 human gene records, whereas T00759, shared by all of our selected genes, has 161 records. T03826 and T00781 do not have named gene records but have some short sequences. Therefore, we focused on T03828 and T00759. We considered the similarity between our selected 7 genes with TRANSFAC and regulated genes of transcription factors in terms of biological process, molecular function, and cellular component, 3 features used by Gene Ontology (GO) to describe genes. In GO, 2 genes or 2 groups of genes with either common functional or location information are considered similar in any of the 3 features. The web-based toolkit WebGestalt is a useful software that incorporates information from different publicly available databases and enables users to find desired patterns[32]. We applied this toolkit by comparing each gene regulated by T03828 and T00759 with the 7 selected genes and performing statistical tests to measure their similarities. We choose the hypergeometric test [32] to calculate the statistics and selected comparable pairs with *P* values less than 0.01. Details of *P* value calculation were previously provided by Zhang *et al.*[32]. The results are listed in Tables 4 and 5.

The first column in both tables lists genes or gene groups that share properties listed in the third columns, with the regulated genes recorded in the second column. Genes set off by semi-colon were analyzed in the same statistical test that showed similarities. In the second column, genes are separated by diagonal marks, and this denotes these genes share properties with candidate genes listed in column 1 based on independent test. In

Table 3. Seven genes with TRANSFAC records and their transcription factors indicated by ChIP-chip fragments

| Gene name | Number of fragments | Transcription factor |
|---|---|---|
| CBX3 | 18 | T00759, T03286, T03828 |
| TES | 6 | T00759, T00781 |
| CALU | 21 | T00759, T00781 |
| HBP1 | 17 | T00759, T00781, T03286, T03828 |
| MDH2 | 12 | T00759, T00781, T03828 |
| PIK3CG | 5 | T00759 |
| NT5C3 | 6 | T00759, T00781, T03828 |

Table 4. **Transcription factor (TF) T03828-regulated genes and similar properties between those genes and the selected gene group**

| Selected gene (s) | TF-regulated genes | Similarities | *P* value of similarity |
|---|---|---|---|
| *MDH2* | *AKR1C4* | Oxidoreductase activity[a] | 1.19E-03 |
| | *AMBP* | Cofactor catabolism[b] | 8.95E-05 |
| | *GK* | Alcohol Metabolism[b] | 5.99E-03 |
| | *GK* | Mitochondrial part[c] | 7.72E-03 |
| *NT5C3;CALU* | *CYP2D6/CYP3A4/CYP8B1* | Endoplasmic reticulum[c] | 3.80E-03 |
| *PIK3CG;MDH2;NT5C3;CALU* | *CYP27A1/CYP2D6* | Cytoplasmic part[c] | 4.62E-03 |

T03828 regulates 20 genes; 7 of 20 have similar properties to our gene set. [a], molecular function; [b], biological process; [c], cellular component.

Table 5. **TF T00759-regulated genes and similar properties between those genes and the selected gene group**

| Selected gene (s) | TF-regulated genes | Similarities | *P* value of similarity |
|---|---|---|---|
| *MDH2* | *ABCA2/ DBH/LDLR/MAOA/MAOB/SOAT1* | Alcohol metabolism[a] | 5.99E-03 |
| | *HSD17B1/HSD3B1/HSD3B2* | Oxidoreductase activity[b] | 1.19E-03 |
| | *GPD2* | Glucose catabolism[a] | 3.39E-04 |
| | *HPSE* | Cellular carbohydrate metabolism[a] | 8.63E-03 |
| | *HSD3B1/HSD3B2* | Mitochondrial part[c] | 7.72E-03 |
| *NT5C3* | *ADORA2A/DHFR* | Nucleotide metabolism[a] | 3.34E-03 |
| | *INPPL1* | Phosphoric monoester hydrolase activity[b] | 6.76E-03 |
| *HBP1* | *BRCA1/EGFR/VHL* | Negative regulation of progression[a] through cell cycle[a] | 2.65E-03 |
| | *CDKN1B/CDKN1A/TP53* | Cell cycle arrest[a] | 4.35E-04 |
| | *CSNK1A1/GPD2* | Wnt receptor signaling pathway[a] | 1.09E-03 |
| *CALU* | *F2R/PSEN1* | Golgi apparatus[c] | 8.73E-03 |
| *CBX3* | *MPO/TOP2B* | Chromatin binding[b] | 5.20E-04 |
| | *PSEN1/PTTG1/TERT* | Chromosome organization and biogenesis[a] | 8.47E-03 |
| | *SOX3/TOP2B* | Chromatin[c] | 3.21E-03 |
| | *TERT* | Chromosomal part[c] | 8.09E-03 |
| *NT5C3;CALU* | *CYP3A4/ CYP3A7/CYP4B1/MGST1 /PSEN1/SOAT1/VHL* | Endoplasmic reticulum[c] | 3.80E-03 |
| *PIK3CG;MDH2; NT5C3;CALU* | *ABCA2/BRCA1/CASP8/CTSL/ CYP27A1/DFFB/EGFR/GPD2/ HPSE/MAOA/ MAOB/MPO/ MYH7/TP53/SPHK1/TOP2B* | Cytoplasmic part[c] | 4.62E-03 |

T00759 regulates 161 genes; 38 of 161 have similar properties to our gene set. [a], biological process; [b], molecular function; [c], cellular component.

other words, the transcription factor co-regulated genes were tested one-by-one with the selected set and are presented together in the tables.

As shown in both Tables 4 and 5, *MDH2* was the most active gene in several similarity measurements. As a mitochondrial precursor, it is included in the process of metabolism and catabolism together with several other genes. *CBX3*, which cooperates in chromosome organization and chromatin locations, was also very active in terms of similarity. Furthermore, we observed that *HBP1* shares its mobility function in cell cycle and receptor pathways with genes regulated by T00759. A group of 4 genes, 3 of which have not been mentioned, showed up with dozens of genes at cytoplasmic parts (Tables 4 and 5, bottom). They were all related to intracellular activities.

## Indirect verification of seven motifs

We next identified shared properties of the genes regulated by T03828 and T00759 and the 7 selected genes. TRANSFAC could only directly verify 21 of 28 candidate motifs with ChIP-chip fragments, leaving 7 unconfirmed. However, we extended the scope wider by incorporating properties shared by co-regulated genes found in previous steps. Because co-regulated genes share genetic sequence compositions, we followed the strategy illustrated in Figure 6 to verify these 7 motifs. If ChIP-chip fragments were unable to directly verify motif *n*, we started from its location genes (for example, gene A in Figure 6) where the SNP is located. We first employed fragments to infer experimentally verified transcription factors, as each fragment in TRANSFAC is mapped to a transcription factor. Then, by previously described similar gene search of the co-regulated gene suggested by transcription factor, we reduced the number of co-regulated genes with similar properties with a statistical test *P* value lower than a threshold. GO information was then incorporated. Then, we used TRANFAC to search motif *n* in the binding fragments or transcription factor database of the chosen co-regulated genes. If motif *n* is traced back to any of these genes, specifically binding fragments, we confirmed that motif *n* was not random noise or false detection.

We selected co-regulated genes with similar properties to *HBP1*, on which 5 of 7 unverified motifs locate, and the four-gene group containing *PIK3CG*, *MDH2*, *NT5C3*, and *CALU*, which has common candidate co-regulated genes. In addition, this group and the co-regulated genes with similar properties to *MDH2* have a number of genes in common (Tables 4 and 5). Thus, although *MDH2* has a larger number of genes in common, we chose HBP1 and the four-gene group to search for binding fragments for the 7 unverified motifs. From these ChIP-chip sequences, we identified motifs 19, 21, 22, 25, 26, and 28, whereas motif 27, which is short in number of location genes, could not be verified. Therefore, in 28 SNP-containing motifs among a group of 10 co-expressed genes, 21 motifs were verified directly with ChIP-chip fragments of their location genes, whereas 6 were confirmed with an indirect strategy using TRANSFAC. Only one motif, CAGGTTCAC, could not be validated with either method. Though this does not necessarily indicate false detection, further experimental validation is still needed.

## Another example of myeloid data analysis

To illustrate the effectiveness of the proposed method, we performed another analysis focusing on myeloid samples only—4 disease and 3 controls. Following the methodology described above, we identified a gene set with 21 genes, illustrated in Table 6. The candidate motifs and their statuses were validated using the TRANSFAC database (Table 7). Based on the results, we concluded that by using samples from the same tissue category, we obtained a larger co-expressed

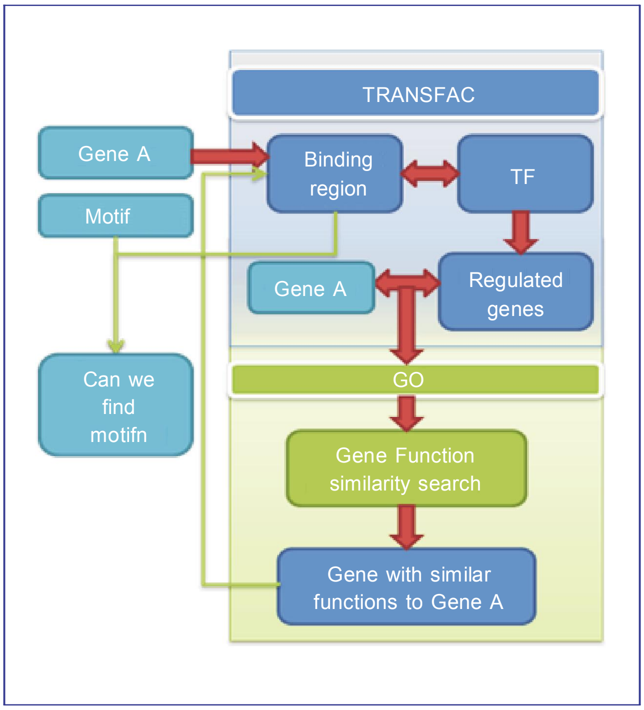

Figure 6. Flowchart of verification of motifs that could not be directly confirmed by TRANSFAC.

gene set as well as a larger candidate TFBS set containing most of the genes and motifs derived using all samples. The results are very consistent except that the pure tissue type tended to include more genes in the co-expressed gene set. As shown in Table 7, all motifs in Table 2 except for GAGTTCCA were identified in the myeloid example. The consistency between different data using the same methodology illustrates the capability of the proposed strategy.

## Discussion

We would like to briefly discuss some methodology and result-related issues in the following paragraphs. The first topic is the application of the proposed association method. As previously suggested, this strategy focuses on genetic variations in TFBS and the regulatory events introduced by these genetic variations. Generally, SNP studies look for bases that have greater than a certain minor allele frequency (e.g., 10% of the population has the minor allele) to show that the SNP is informative. On the other hand, mutation can also occur as a one-off event at few individual sites. Therefore, SNPs can be considered a special case similar to genetic mutations that occur in a population. Moreover, mutation-related disease rates are far lower than the threshold that we normally use to define SNPs (10%); in mutation-related disease studies, the genetic variations are actually mutations. Thus, the proposed strategy could be applied to studies assuming genetic variations, including mutations and SNPs, that are highly involved in the disease mechanism and progression.

Many feature selection methods can be applied for motif selection, including support vector regression with recursive feature elimination, Bayesian methods, and piecewise linear networks. All of these approaches can be adopted in our strategy. Stepwise regression is the most popular form of feature selection in statistics. LASSO and its generalization, LARS-EN, use cross validation to identify the best size of subgroups obtained by these greedy algorithms. Considering the complexity of computation and time required for the previously mentioned feature selection and subset identification methods, as well as their unproven superiority to LARS-EN in solving the motif selection problem, LARS-EN is a

Table 6. Co-expressed gene cluster from myeloid samples

| Gene names | | | | | | |
|---|---|---|---|---|---|---|
| *MCA2_HUMAN* | *EIF2AK1* | *PSCD3* | *SCIN* | *CBX3* | *DNAH11* | *TES* |
| *PON1* | *CALU* | *KLHL7* | *HBP1* | *HIPK2* | *MDH2* | *OSBPL3* |
| *WNT2* | *TAX1BP1* | *PIK3CG* | *NT5C3* | *JAZF1* | *CDK6* | *BPGM* |

Table 7. Motifs identified from myeloid samples

| Motif | Status | Motif | Status | Motif | Status |
|---|---|---|---|---|---|
| AATGGG | Direct | CTGTCACT | Direct | TTTTGGAGT | Indirect |
| CATTGC | Direct | GAGTTCCA | Direct | AAAAAGAA | Direct |
| GGACTC | Direct | TTTTGGAG | Direct | CCCTGCAGA | Indirect |
| TCAGGG | Direct | ATAAATGA | Direct | GGTATGTTG | Direct |
| CACTTT | Direct | AGGAAAAT | Direct | CTTGCTGCC | Direct |
| TGCAGAT | Indirect | GGCAGATT | Direct | AGGCAGATT | Direct |
| ATTTCTC | Direct | CTTAAAAT | Direct | CAGGTTCAC | Unverified |
| TTTCACT | Direct | CATGTGAA | Direct | CATTTGGTA | Indirect |
| TTCACTT | Direct | CTTGGATA | Indirect | CTCATTTGAC | Indirect |
| GTGCCAC | Direct | TGTGCCAC | Indirect | ATCTCCTGCC | Indirect |
| CTGTGTCA | Direct | TGGGAGCG | Direct | ATTTCATTTT | Indirect |
| TTTAGAAA | Direct | AACACCTC | Indirect | CCTGCCTCC | Indirect |
| TATGTTAT | Indirect | GAGCAACC | indirect | | |

Status indicates whether the motif can be directly or indirectly verified by TRANSFAC, or unverified.

very suitable strategy that is simple and takes advantage of simultaneous shrinkage and model selection to track the group effect.

As mentioned above, one cluster comprised of 10 genes was investigated to obtain the motifs and derive conclusions. However, there were other clusters that fit the predefined conditions of SNP number and copy number variations. We used the selected genes as an example to illustrate our methodology, and all results were obtained based on this group. Identification of a complete list of candidate genes and TFBS motifs to select important biomarkers will require additional studies. For example, the same strategy could be performed on all co-expressed clusters genome-wide, followed by selection candidate biomarkers and validation experiments. In addition, possible TFBS identified in this work are based on a cluster of genes and concept of motifs. Therefore, they are shared events in a predefined set. In addition, regulatory candidates for individual genes may be filtered in the motif selection process. Therefore, our strategies are not guaranteed to identify all the TFBS candidates, especially binding events for a single gene, because the system is designed based on co-expressed gene clusters.

## Conclusions

The frequency and incidence of MDS are increasing in the American population, which has an estimated annual incidence of about 3.5 to 10 per 100,000 in the general population and 12 to 50 per 100,000 in the elderly population[33]. However, the prognosis of MDS patients has not shown any significant improvement over the last decade[34]. Therefore, the success of identifying biomarkers for accurate diagnosis and prognostic stratification of MDS will eventually lead to significant improvement in patient outcome. In this study, we associated SNP array data and gene expression array data to evaluate genes with genetic variations and mutations that may be the causal mechanism for different RNA expression profiles. In particular, we identified candidate motifs introduced by genetic variations and mutations in transcription binding events, which could be considered biomarkers for the disease, with additional information. To do this, differentially expressed genes with copy number gain or loss were selected and clustered. Then we concentrated on a co-expressed group and introduced SNP-containing motifs as possible TFBS. Based on the assumption that only a specific genetic variation would introduce certain biological events (e.g., binding of a special regulatory transcription factor may alter gene expression and drive disease pathogenesis), we identified 28 SNP-containing motifs in the selected gene group. The TRANSFAC database, which is a collection of experimentally verified transcription factor and binding sequences, was used to verify the 28 motifs. ChIP-chip fragments verified 21 motifs directly. Then we studied genes co-regulated by the same transcription factor as indicated by ChIP-chip fragments and their functional similarities. With co-regulation and functional similarity, the selected gene group and the co-regulated gene group were employed to infer proofs for the 7 unverified motifs. This step verified 6 of 7 motifs that could not be directly located in TRANSFAC. These selected genes and candidate TFBS integrated copy number variation and genotyping information to complete a list of candidate biomarkers to be further tested experimentally. To summarize, the method we suggest serves as a linkage between two types of array profiles of the same disease model and identifies regulatory motifs introduced by genetic variations. Our results suggest SNP-containing motifs as TFBS may be another direction for mechanistic study and biomarker discovery for MDS.

## Acknowledgment

The author would like to thank Dr. Ling-Yun Wu with Institute of Applied Mathematics, Chinese Academy of Science and Dr. Xiao-Rong Yang with Zhejiang Gongshang University for their suggestions and colleagues at The Methodist Hospital Research Institute, for their discussion in the research. This work was partially supported by grants from NIH (No. 1R01LM010185, 1U01CA166886, and 1U01HL111560).

Received: 2012-05-02; revised: 2012-11-05;
accepted: 2012-11-15.

## References

[1] Gondek LP, Dunbar AJ, Szpurka H, et al. SNP array karyotyping allows for the detection of uniparental disomy and cryptic chromosomal abnormalities in MDS/MPD-U and MPD. PLoS ONE, 2007,2:e1225.

[2] Gondek LP, Tiu R, O'Keefe CL, et al. Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. Blood, 2008,111:1534-1542.

[3] Chen G, Zeng W, Miyazato A, et al. Distinctive gene expression profiles of CD34 cells from patients with myelodysplastic syndrome characterized by specific chromo-somal abnormalities. Blood, 2004,104:4210-4218.

[4] Pellagatti A, Esoof N, Watkins F, et al. Gene expression profiling in the myelodysplastic syndromes using cDNA microarray technology. Br J Haematol, 2004,125:576-583.

[5] Lastowska M, Viprey V, Santibanez-Koref M, et al. Identification of candidate genes involved in neuroblastoma progression by combining genomic and expression microarrays with survival data. Oncogene, 2007,26:7432–7444.

[6] Lum PY, Chen Y, Zhu J, et al. Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. J Neurochem, 2006,97 Suppl 1:50–62.

[7] Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol, 2008,6:e107.

[8] Virtanen IM, Noponen N, Barral S, et al. Putative susceptibility locus on chromosome 21q for lumbar disc disease (LDD) in the finnish population. J Bone Miner Res, 2007,22:701–707.

[9] Yamazaki K, Onouchi Y, Takazoe M, et al. Association analysis of genetic variants in IL23R, ATG16L1 and 5p13.1 loci with Crohn's disease in Japanese patients. J Hum Genet, 2007,52: 575–583.

[10] Milet J, Dehais V, Bourgain C, et al. Common variants in the BMP2, BMP4, and HJV genes of the hepcidin regulation pathway modulate HFE hemochromatosis penetrance. Am J Hum Genet, 2007,81:799–807.

[11] Guenard F, Labrie Y, Ouellette G, et al. Mutational analysis of the breast cancer susceptibility gene BRIP1 /BACH1/FANCJ in high-risk non-BRCA1/BRCA2 breast cancer families. J Human Genet, 2008,53:579–591.

[12] Wang M, Vikis HG, Wang Y, et al. Identification of a novel tumor suppressor gene p34 on human chromosome 6q25.1. Cancer Res, 2007,67:93–99.

[13] Kolbehdari D, Wang Z, Grant JR, et al. A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. J Dairy Sci, 2008,91:2844–2856.

[14] Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science (New York, USA), 2007,316:1331–1336.

[15] Yang HH, Hu N, Taylor PR, et al. Whole genome-wide association study using affymetrix SNP chip: a two-stage sequential selection method to identify genes that increase the risk of developing complex diseases. Methods Mol Med, 2008,141:23–35.

[16] Delforge M. Understanding the pathogenesis of myelodysplastic syndromes. Hematol J, 2003,4:303–309.

[17] Ylipää A, Yli-Harja O, Zhang W, et al. A systems biological approach to identify key transcription factors and their genomic neighborhoods in human sarcomas. Chin J Cancer, 2011,30: 27.

[18] Honeycutt E, Gibson G. Use of regression methods to identify motifs that modulate germline transcription in Drosophila melanogaster. Genet Res, 2004,83:177–188.

[19] Keles S, van der Laan MJ, Vulpe C. Regulatory motif finding by logic regression. Bioinformatics (Oxford, England), 2004,20: 2799–2811.

[20] Liu KY, Zhou X, Kan K, et al. Bayesian variable selection for gene expression modeling with regulatory motif binding sites in neuroinflammatory events. Neuroinformatics, 2006,4:95–117.

[21] Liu G, Zhang W. Will chinese ovarian cancer patients benefit from knowing the BRCA2 mutation status? Chin J Cancer, 2012,31:1.

[22] Ma H, Zhou Z, Wei S, et al. Association between p21 Ser31Arg polymorphism and cancer risk: a meta-analysis. Chin J Cancer, 2011,30:254.

[23] Zhou C, Wang J, Cao S, et al. Association between single nucleotide polymorphisms on chromosome 17q and the risk of prostate cancer in a Chinese population. Chin J cancer, 2011,30:721.

[24] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc: Series B Stat Methodol, 2005,67: 301–320.

[25] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol, 1996,58:267–288.

[26] Matys V, Fricke E, Geffers R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res, 2003,31:374–378.

[27] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England), 2003,4:249–264.

[28] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, 2001,98:5116–5121.

[29] de Hoon MJ, Imoto S, Nolan J, et al. Open source clustering software. Bioinformatics (Oxford, England), 2004,20:1453–1454.

[30] Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci, 2003,18:71–103.

[31] Birzele F, Kramer S. A new representation for protein secondary structure prediction based on frequent patterns. Bioinformatics (Oxford, England), 2006,22:2628–2634.

[32] Zhang B, Kirov S, Snoddy J. Webgestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res, 2005,33:W741–W748.

[33] Aul C, Bowen DT, Yoshida Y. Pathogenesis, etiology and epidemiology of myelodysplastic syndromes. Haematologica, 1998,83:71–86.

[34] Kasper DL, Braunwald E, Fauci AS, et al. Harrison's principles of internal medicine. 16th Ed. New York, USA: McGraw-Hill Companies, 2005.