Genomics 98 (2011) 64–71

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

Differential expression pattern-based prioritization of candidate genes through integrating disease-specific expression data

Yun Xiao [a,1], Chaohan Xu [a,1], Yanyan Ping [a], Jinxia Guan [a], Huihui Fan [a], Yiqun Li [a], Xia Li [a,b,*,1]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150086, China
[b] Department of Bioinformatics, Capital University of Medical Sciences, Beijing 100084, China

ARTICLE INFO

Article history:
Received 17 August 2010
Accepted 1 April 2011
Available online 15 April 2011

Keywords:
Prioritization
Differential expression pattern
Complex disease
Integration

ABSTRACT

Expression data can reveal subtle transcriptional changes that mediate the clinical phenotype of the disease resulting from interaction between genetic and environmental factors, which offers us a new perspective to prioritize candidate genes. Here, we proposed a novel differential expression pattern (DEP)-based approach integrating numerous disease-specific expression data sets for prioritizing candidate genes. Using breast cancer as a case study, we validated the efficiency of our approach through integrating 12 breast cancer-related expression data sets based on the leave-one-out cross-validation. Particularly, prioritization based on subtype-specific expression data sets could generate significantly higher performance. The performance could be continually improved with the increasing expression data sets regardless of platform heterogeneity. We further validated the robustness of this approach by application to prostate cancer. Additionally, our approach showed higher performance in comparison with other expression-based approaches and better capability of identification of less well-studied disease genes in comparison with other integration-based approaches.

1. Introduction

Identification of disease genes plays critical roles in understanding the pathogenesis of complex disease (e.g., cancer, heart disorder, and diabetes) and developing new drugs for prevention and treatment of complex disease. In the past two decades, a large number of linkage analysis and association studies, which produce several hundreds of candidate genes, have been performed. However, the experimental evaluation of the complete list of candidate genes is time-consuming and expensive. Hence, many bioinformatics approaches for prioritization of candidate genes have been developed to assist identification of disease genes.

Previous prioritizing methods use a variety of biological data, such as sequence [1], function [2–6], expression [7–11], network [12–15], text-mining [16–18], or combinations of them [19–26]. For example, PROSPECTR [27] ranked candidate genes based on various sequence features, which show a significant difference between known human disease genes and genes not known to be involved in disease. Using protein–protein interaction networks, Kohler et al. [28] proposed a method that prioritizes candidate genes by use of a random walk-based similarity measure. ENDEAVOUR [29] fused multiple biological data to prioritize candidate genes by an order statistics-based computational method.

Notably, among these biological data, expression data show the most rapid increase in the past ten years due to the advancement of all kinds of high-throughput biological technologies, such as microarray and next-generation sequencing. Indeed, most previous prioritization approaches used expression data [1,2,6,14,24,29–32] to calculate Pearson correlation coefficients between genes for approximately quantifying potential functional relationships and prioritized candidate genes by using these co-expression relationships based on the hypothesis that disease genes tent to exhibit similar functions. However, only single expression data were utilized (e.g., the human atlas expression data [33]). Recently, many approaches based on multiple expression data sets were developed for prioritization of candidate genes. For example, TOM [34] extracted gene co-expression relationships in a large number of expression data sets derived from different anatomic sites and tissues, different platforms and different conditions to prioritize candidate genes. Piro et al. [35] described a candidate gene prioritization approach based on the spatial gene-expression patterns generated by combining multiple 3D expression data from an entire organ. Oti et al. [36] calculated evolutionary conservation co-expression scores by integrating multiple non-disease-specific expression data from five distinct species including yeast, worm, fly, mouse, and human to prioritize candidate genes. However, Oti et al. [36] observed that the performance is dependent on the expression data used, and cannot be improved when combining more expression data sets.

Analyzing genome-wide transcriptional changes has been performed in a wide range of human diseases, which can effectively capture the intermediate response of disease gene-induced phenotypes. Therefore, transcriptional changes offer a possibility for identifying disease genes. A

* Corresponding author at: College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China. Fax: +86 451 86615922.
E-mail address: lixia@hrbmu.edu.cn (X. Li).
[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2011.04.001

recent work designed by Chen et al. [37] demonstrated that highly differentially expressed genes are more likely to be disease genes, further supporting the potential relationship between disease genes and transcriptional changes. However, due to the influence of disease heterogeneity, population difference, and environmental factors on gene expression, the relationship can be unclear in the case of individual expression studies. We postulated that transcriptional change patterns (also termed differential expression patterns, DEPs) across a large number of disease-associated expression studies can more comprehensively and accurately reflect this potential relationship, which can be used for assisting identification of disease genes.

In our study, we proposed a novel DEP-based prioritization approach (Fig. 1) by integrating many disease-specific expression data sets. Using breast cancer as a case study, we evaluated the performance of our DEP-based prioritization approach by the leave-one-out cross-validation. Our results showed that the DEP-based approach can effectively prioritize candidate genes, especially using subtype-specific expression data sets. Several factors possibly influencing the performance were also investigated, such as the number of expression data sets, platform heterogeneity, and the number of known disease genes. We found that the performance can be continually improved with the increase of expression data sets used. We also compared with other expression-based, sequence-based, and integration-based candidate gene prioritization approaches. Comparison results showed that the DEP-based prioritization approach has better performance in comparison with other expression-based and sequence-based prioritization approaches. Comparisons with integration-based prioritization approaches suggested that our approach can effectively identify less well-studied disease genes.

2. Methods

2.1. Expression and Disease Gene Data

Disease-associated case-control expression data sets were obtained from the GEO and ArrayExpress databases. In order to get more reliable differential expression results, expression data sets with the number of disease or normal samples less than four were removed. Each expression data set was imputed using the k-nearest-neighbor method, and was then normalized using the median normalization method. All gene expression values were log2-transformed. Known disease genes were obtained from the Online Mendelian Inheritance in Man (OMIM) database [38], and their corresponding chromosomal regions were obtained from the Entrez Gene database.

2.2. DEP-based Prioritization Approach

2.2.1. Construction of the DEP Matrix

Given a specific disease which contains k known disease genes, N case-control expression data sets were obtained. For each expression data set, differentially expressed genes were determined using the Statistical Analysis of Microarrays (SAM) algorithm [39] with a significant level of 10% ($p<0.1$). M genes consistently present in all of these expression data sets were defined as background genes. For a given gene i in the background genes, a differential expression binary vector $x_{i1},\ldots,x_{ij},\ldots,x_{iN}$ was produced according to its expression changes across the N case-control expression data sets, where x_{ij} indicates whether the gene i shows significantly differential expression in the jth expression data set. If the gene i was identified as a differentially expressed gene in the jth

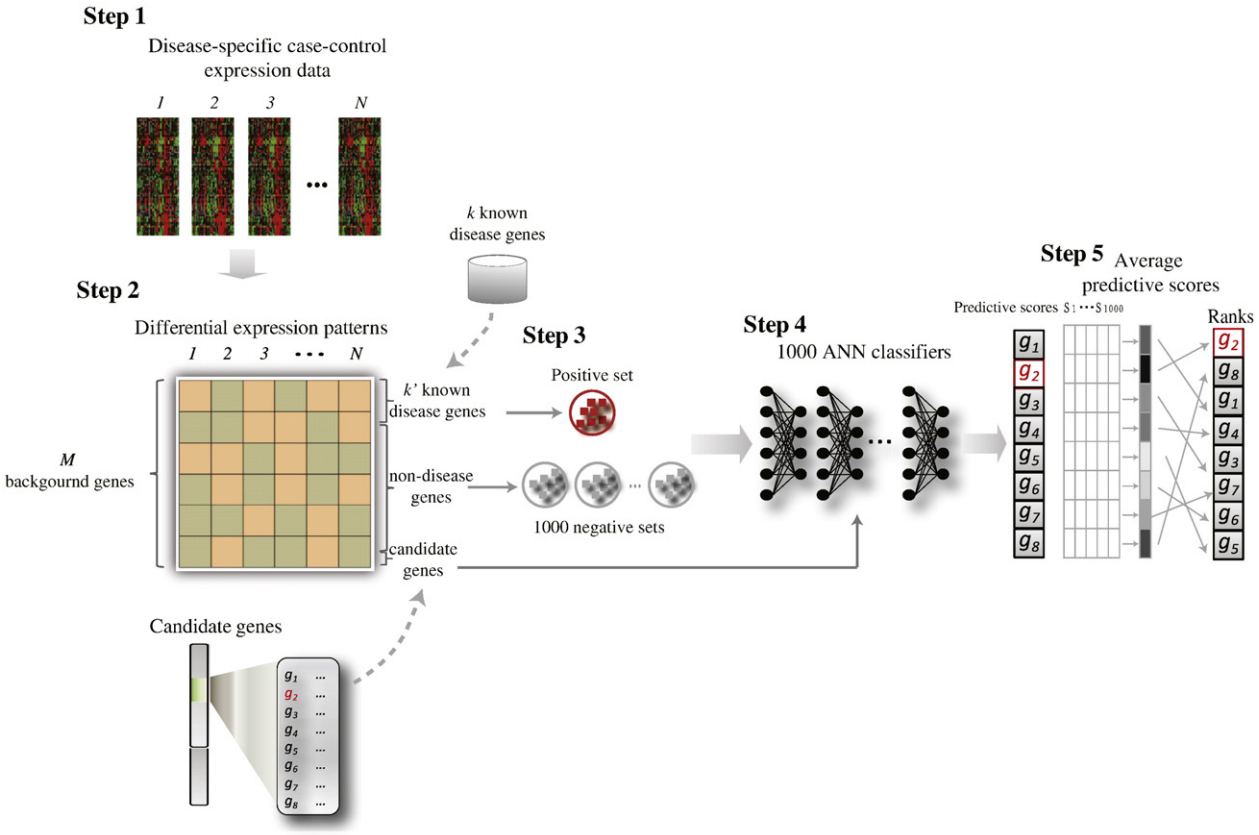

Fig. 1. Workflow of the DEP-based prioritization approach. Candidate genes were prioritized based on their DEPs across numerous disease-specific expression data in a five-step analysis. Step 1, multiple disease-specific case-control expression data sets were gathered from the GEO and ArrayExpress databases and known disease genes were derived from the OMIM database. Step 2, the DEP matrix was constructed on the basis of the occurrences of differentially expressed genes in these expression data sets. Step 3, construction of training sets using known disease genes and randomly selected non-disease genes. Step 4, 1000 ANN classifiers were built according to the DEPs. Step 5, candidate genes were predicted by these ANN classifiers, and the average predictive scores were calculated, which were used for ranking.

expression data set, then $x_{ij}=1$, otherwise $x_{ij}=0$. Then, a $M\times N$ DEP matrix was generated, in which each row represents a gene and each column represents a case-control expression data set.

2.2.2. Construction of ANN Classifiers and Prioritization of Candidate Genes

Given a candidate gene set (e.g., positional candidate genes from linkage analysis), we constructed artificial neural network (ANN) classifiers based on the DEP matrix to rank these candidate genes.

Let k' be the number of known disease genes with expression values in all of the N expression data sets and significantly differential expression in at least one of these data sets. These genes were regarded as the positive set for training the prioritization model. In order to generate the negative set, we constructed an artificial non-disease gene set that refers to all background genes excluding disease genes and candidate genes. Due to the imbalance of the numbers of disease and non-disease genes, we randomly selected k' non-disease genes as the negative set. Based on the DEPs of the positive and negative sets, a typical ANN classifier was trained for assigning scores to candidate genes. The scores were used for representing the probability of candidate genes identified as disease genes.

In order to avoid the overfitting of the ANN classifier trained, 1000 negative gene sets randomly selected from the artificial non-disease gene set were generated, which, together with the same positive set, were used to construct 1000 ANN classifiers. All candidate genes were assigned with predictive scores using each ANN classifier. Subsequently, for each candidate gene, an average predictive score $\overline{S}$ across these 1000 classifiers was calculated:

$$\overline{S}=\frac{\sum_{l=1}^{1000} s_l}{1000}$$

where s_l is the predictive score of this gene in the *lth* ANN classifier. Finally, candidate genes were ranked according to the average predictive scores in a descending order.

The ANN contained three layers: an input layer, a hidden layer and an output layer. Every neuron in the hidden layer and output layer was weightedly connected with all neurons in its previous layer and activated by the tan-sigmoid transfer function. The number of input neurons was set to N and the number of neurons in the hidden layer was set to 20. The output layer contained only a single neuron, whose output was a score ranged from 0 to 1. The training of ANN classifiers and prediction were carried out by using the MATLAB Neural Network Toolbox (the number of training cycles was set to 200 and the specified error goal was 1e−5).

2.3. Cross-validation

The performance of the DEP-based prioritization approach was evaluated by the leave-one-out cross-validation. For each disease gene, we constructed an artificial genetic interval of 20 Mb centered on this disease gene because the resolution of traditional genetic linkage analyses was usually restricted to 10 to 30 cM [40]. Those genes within the artificial locus for which all expression data are available were regarded as the test set, and the remaining disease genes were denoted as the positive training set. Genes not belonging to the positive and test sets were used to create 1000 negative training sets. Then, we exploited the positive and negative sets to train the prioritization models for ranking the test genes. The ranks of the test genes were transformed into relative ranks by:

$$R_i=\frac{(n-r_i)}{(n-1)}$$

where n is the total number of the test genes and r_i represents the rank of a test gene i in the test set. The relative rank of the test gene at the top of the rank list was set to 1.0, and the relative rank of the test gene at the bottom was set to 0.0. Then, the relative rank of the true disease gene was recorded.

The receiver operating characteristic (ROC) curve was used to measure the performance of the DEP-based prioritization approach. It was plotted as 1-specificity (i.e., false positive rate) versus sensitivity for all thresholds in the range of relative ranks of disease genes. The area under the ROC curve (AUC) was used as a standard measure for evaluating the overall prediction performance of our approach. For instance, an AUC value of 100% suggests that every disease gene was ranked at the top of the corresponding test gene list.

2.4. Evaluation in Breast Cancer Using Original and Subtype-Specific Expression Data Sets

Initially, we used original breast cancer expression data sets to validate our prioritization model. Because breast cancer is generally considered as a complex disease characterized by various intrinsic subtypes which show genetic differences in the pathogenic mechanisms, we constructed subtype-specific expression data sets using these original data sets. For each original expression data set, its breast cancer samples were divided into subtypes by using the strategy described in [41], and then breast cancer samples with different subtypes respectively together with normal samples were used for forming new subtype-specific expression data sets. These subtype-specific expression data sets with less than four disease samples were removed. Using these subtype-specific expression data sets, we re-evaluated the performance of the DEP-based prioritization approach.

2.5. The Influence Factors of the DEP-based Prioritization Approach

Several influence factors including the number of expression data sets, sample size, the heterogeneity of microarray platforms, and the number of disease genes as well as the number of ANN classifiers were evaluated independently. For each influence factor, we constructed different DEP-based prioritization models by changing this factor and assessed the effect of this factor on the performance. In order to assess the effect of the number of expression data sets, many different DEP-based prioritization models were created based on different numbers of expression data sets. To explore the effect of sample size, we randomly extracted different numbers of samples from a specific expression data set, which combined with the other data sets, were used to construct multiple prioritization models. As for the platform heterogeneity, prioritization models were constructed using the same number of expression data sets from mixed platforms and consistent platforms. Different numbers of disease genes and ANN classifiers were, respectively, used to construct prioritization models to detect the influence of the numbers of known disease genes and ANN classifiers. Details were supplied in the Supplementary Methods.

2.6. Comparisons with Other Prioritization Methods

We compared our approach with several expression-based methods (co-expression, conserved co-expression [36] and differential expression ratio [37]) and sequence-based PROSPECTR as well as integration-based ENDEAVOUR. In order to compare with these methods, we performed the leave-one-out cross-validation. For each disease gene, genes located at the 20 Mb region around this disease gene were selected as the test set, and the rest of the disease genes were called the training set. These methods were used to rank the test genes based on the training set. Finally, we plotted ROC curves and computed AUC scores to assess the performance of these methods.

The programs of our approach are available at http://bioinfo.hrbmu.edu.cn/DEP/DEP.html. Detailed methods for comparisons were provided in Supplementary Methods.

3. Results

3.1. Evaluation of the DEP-based Prioritization Approach in Breast Cancer

Through an extensive collection of breast cancer associated expression data sets from the GEO and ArrayExpress databases, 12 breast cancer expression data sets derived from 12 distinct studies were obtained (Supplementary Table S1). These expression data sets included 908 samples (817 breast cancer and 91 normal samples), and a total of 11,633 background genes present in these expression data sets. In these background genes, 31 were referred to as known breast cancer genes in the OMIM database (Supplementary Table S2).

We used the leave-one-out cross-validation to evaluate the performance of our DEP-based prioritization approach based on the 12 original breast cancer expression data sets. Fig. 2A shows the distribution of relative ranks of all known breast cancer genes, which displays a right-leaning trend. There were 64.5% of known breast cancer genes with relative ranks>0.5. Especially, 32.2% were at 0.8 to 1.0 (Supplementary Table S2).

It should be noted that breast cancer with different molecular subtypes can be regarded as separable diseases [41]. Therefore, we suspected that DEPs constructed by subtype-specific expression data sets can further improve the performance of our approach. Based on expression patterns of 534 breast cancer "intrinsic genes", we classified 817 breast cancer samples into five subtypes using the PAM algorithm. Thirty-four subtype-specific expression data sets were generated by recombining breast cancer subtype samples with corresponding normal samples. After filtering subtype-specific expression data sets with less than four disease samples, 30 subtype-specific expression data sets containing 808 disease samples (98 basal-like, 245 ERBB2, 447 luminal-A, 5 luminal-B and 13 normal-like; Supplementary Table S3) were used to evaluate the performance. As shown in Fig. 2B, the distribution of relative ranks shows a stronger right-leaning trend. About 60% relative ranks of known breast cancer genes were at 0.8 to 1.0, and only 12.9% were less than 0.5 (Supplementary Table S2). Fig. 2C shows the ROC curves for original and subtype-specific expression data sets. As expected, results of prioritization using subtype-specific expression data sets were significantly better than those using original data sets—the relative ranks of many breast cancer genes showed an obvious increase and the AUC score increased from 62.2% to 74.1%, suggesting that DEPs constructed using subtype-specific expression data sets can further enhance the power of our prioritization approach.

In addition, we investigated whether genes not known to be involved in breast cancer can be used to prioritize disease genes. We assembled 100 sets of 30 randomly selected genes as positive sets, and then calculated 100 average predictive scores for each test gene using

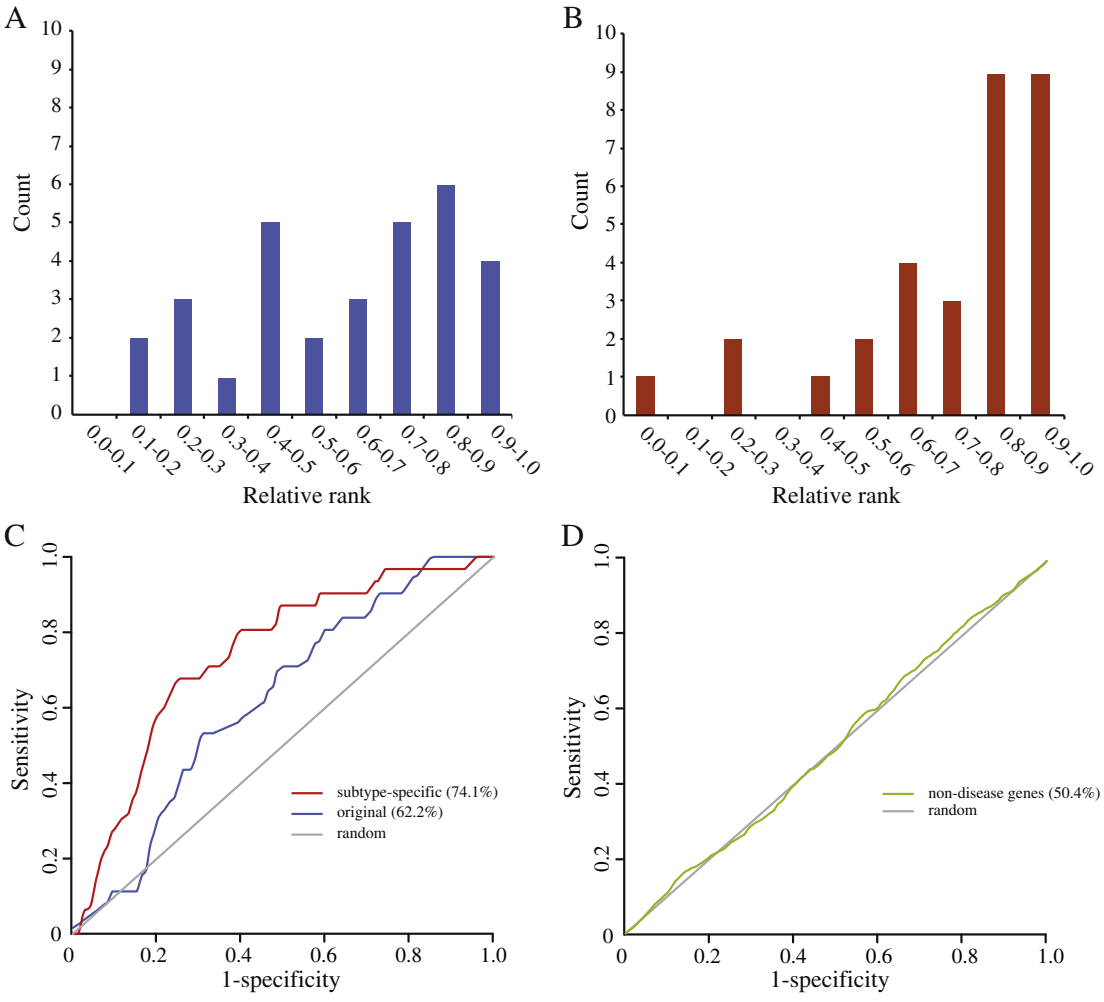

Fig. 2. Evaluation of the DEP-based prioritization approach in breast cancer. A) The histogram shows the distributions of relative ranks of all 31 known breast cancer genes generated using original expression data sets. B) The histogram shows the distributions of relative ranks of all 31 known breast cancer genes generated using subtype-specific expression data sets. C) ROC curves of relative ranks were obtained using original and subtype-specific expression data sets, which are colored with blue and red lines, respectively. D) The ROC curve (green line) was obtained by randomly constructing positive gene sets.

these positive sets. Subsequently, we ranked the test genes according to the mean of the average predictive scores. Fig. 2D shows the ROC curve for randomly selected positive sets. The AUC score was 50.4%, approximating to the random case. Obviously, our approach reached a higher AUC score using disease genes than that using randomly selected genes, indicating that the DEP-based prioritization approach can be sensitive and specific in prioritizing disease genes.

3.2. The Influence Factors of the DEP-based Prioritization Approach

3.2.1. The Number of Expression Data Sets

In order to evaluate whether the number of expression data can affect the performance of our approach, different numbers of expression data sets that were randomly selected from the 12 original expression data sets were used to construct prioritization models (Supplementary Methods). With the increase of the number of expression data, the relative ranks were obviously raised in the applications of both original and subtype-specific data (Fig. 3A and B). For original data sets, the average relative ranks ranged from 0.36 to 0.62 when the number of expression data increased from 2 to 11. Using subtype-specific data sets, the average relative ranks increased from 0.59 to 0.74. The results strongly suggested that increasing expression data sets can improve the performance of our approach. Also, it supported the finding that disease subtype can bring great improvement even with the application of few expression data sets.

3.2.2. Sample Size

The large range of sample sizes across different studies may influence the performance of our approach. The smallest sample size in the 12 original expression data sets is 22 (GSE8977), and the largest sample size is 196 (GSE5346). To detect the influence of sample size, we generated multiple prioritization models with different numbers of samples for a specific data set and then calculated corresponding average relative ranks. By analyzing three expression data sets with the largest sample sizes (GSE5364, GSE9309, and GSE3165), we found that

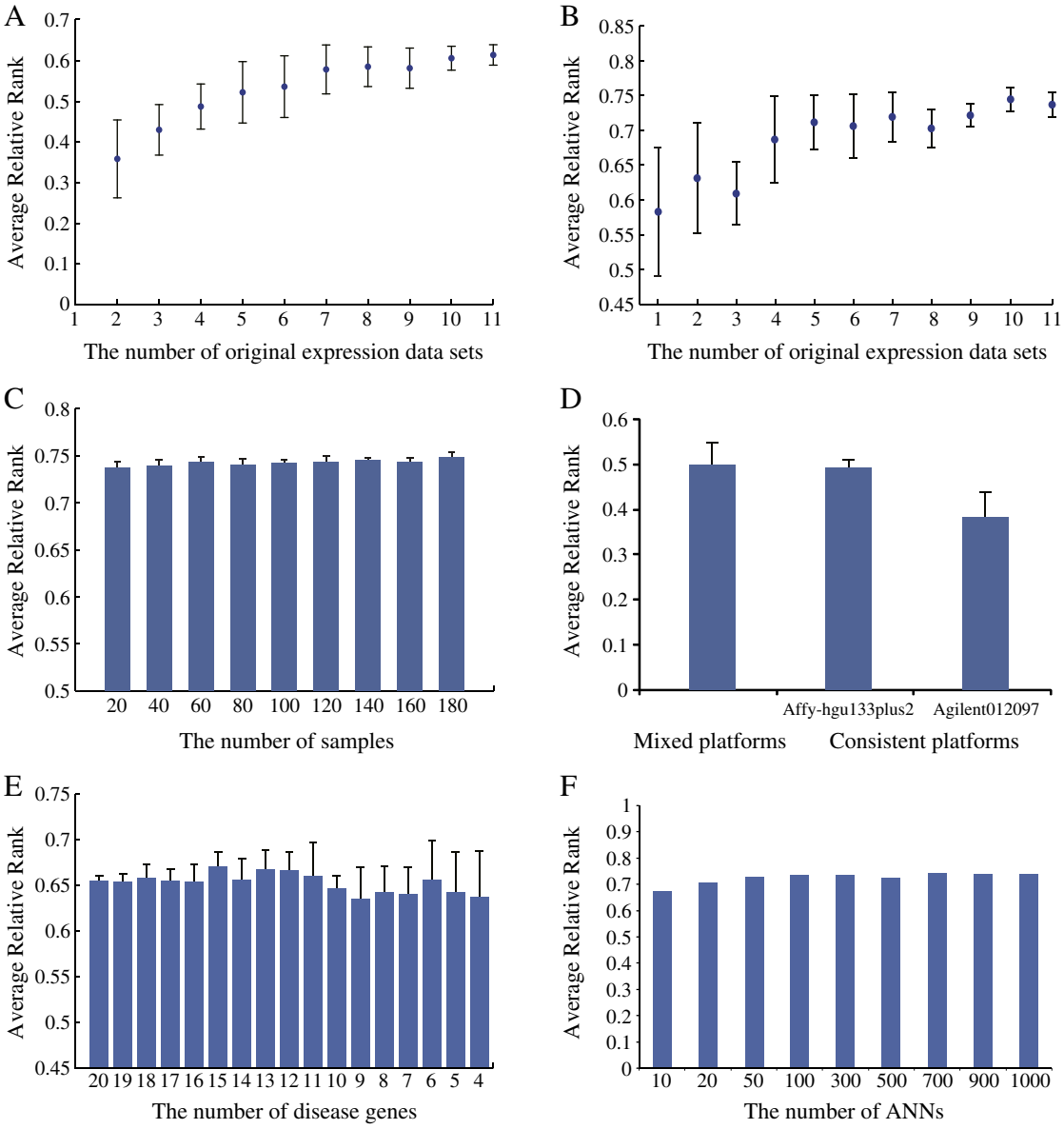

Fig. 3. The effects of multiple factors on the DEP-based prioritization approach. A) The relationship between average relative ranks and the number of original expression data sets. B) Average relative ranks calculated using subtype-specific expression data sets derived from different numbers of original expression data sets. C) The effect of sample size. Newly formed expression data sets with different numbers of samples extracted from the data set GSE5364, together with the other 11 expression data sets, were used to build prioritization models. D) The influence of platform heterogeneity. E) Prioritization using different numbers of known disease genes. F) The influence of the number of ANN classifiers. Error bars represent one standard deviation.

the average relative ranks have a slight increase with the increase of sample size (Fig. 3C and Supplementary Fig. S1). A possible explanation is that the increase of sample size can offer more accurate transcriptional changes between case and control subjects.

3.2.3. The Heterogeneity of Microarray Platforms

Despite the rapid increase of expression data, the heterogeneity of platforms used in different microarray studies might greatly impede the integration of a large number of currently available expression data sets. The 12 original breast cancer associated expression data sets were detected by 4 distinct microarray platforms. To investigate whether the heterogeneity of platforms could affect the performance of the DEP-based prioritization approach, we compared results calculated by randomly selecting four expression data sets with mixed and consistent platforms (Supplementary Methods). Fig. 3D shows that the average relative rank corresponding to the mixed platforms was slightly higher than those from the consistent platforms ($p = 0.90$ for Affymetrix chip and $p = 0.07$ for Agilent chip, t-test). This suggested that prioritization using expression data sets from mixed platforms provides comparable results with consistent platforms, that is, the heterogeneity of microarray platforms has little influence on the performance of our approach, which may be attributed to a recent observation that both Affymetrix and Agilent chips display high consistence with PCR and TaqMan [42]. We also found that the average relative rank calculated by Affymetrix chip is slightly higher than that from Agilent chip, but without statistical significance ($p = 0.14$, t-test).

3.2.4. The Number of Disease Genes

A large number of disease genes with modest effect on complex diseases were not identified yet. We therefore suspected whether diseases with a few disease genes identified can be suitable for prioritization using this approach. We investigated the influence of the number of known disease genes on the prioritization results (Supplementary Methods). Eleven known breast cancer genes were randomly selected as the test genes. From the remaining 20 genes, different numbers of disease genes were randomly selected as the positive sets to train ANN classifiers, and then the relative ranks for the 11 test genes were calculated. As shown in Fig. 3E, the average relative rank changed slightly when the number of disease genes varied from 20 to 4, and the standard deviation increased as the number of disease genes were reduced. We repeated the above process 10 times and obtained similar results (Supplementary Fig. S2), suggesting that our approach was robust to the number of known disease genes and can be effectively used for complex diseases even with a few known disease genes.

3.2.5. The Number of ANN Classifiers

In order to evaluate the effect of the number of ANN classifiers on our approach, we built multiple prioritization models with different numbers of ANN classifiers. We observed that the average relative ranks show a slight increase when the number of ANN classifiers varies from 10 to 50 (Fig. 3F). When the number of ANN classifiers is greater than 50, the average relative ranks reached a steady state (approximately 0.74). Together, our findings suggested that the number of ANN classifiers has minimal effect on our approach.

3.3. Comparisons with Other Candidate Gene Prioritization Approaches

Expression data have been comprehensively used for prioritization of candidate genes. Most of the expression-based prioritization methods calculated co-expression relationships among genes using different expression data to prioritize candidate genes. When compared with the co-expression approach, we found that our approach significantly outperforms the co-expression approach based on both the human atlas expression data and the merged breast cancer expression data set (Fig. 4A). Interestingly, the performance based on multiple breast cancer expression data sets is significantly lower than that based on the atlas expression data, which indicated that the performance of the co-expression method may be dependent on the expression data sets used. Recently, a conserved co-expression prioritization approach integrating multiple non-disease-specific expression data from five species was developed [36]. As shown in Fig. 4A, the DEP-based approach was superior to it (AUC score: 74.1% versus 63.5%). In addition, we also compared our approach with a DER-based method [37] that calculated the ratio of the count of differential data sets to the count of all data sets. Obviously, our approach had a higher AUC score than the DER-based approach using the 12 original and 30 subtype-specific expression data sets (AUC scores: 65.8% for original expression data sets and 69.9% for subtype-specific expression data sets).

Besides expression data, other biological resources were also used for prioritization, such as sequence information. We compared our approach with PROSPECTR [27]. Obviously, our approach was significantly superior to PROSPECTR (Fig. 4B). When compared with integration-based ENDEAVOUR, we found that ENDEAVOUR exhibited an overall better performance than our approach (Fig. 4B). Noticeably, in the four disease genes (*DIRAS3*, *LSP1*, *RB1CC1*, and *SLC22A18*) associated with the fewest PubMed IDs, two had higher ranks in our approach than those in ENDEAVOUR, one had the same ranks, and one was ranked slightly lower in our approach (Supplementary Table S4). Also, we used our approach and ENDEAVOUR to prioritize two recently identified breast cancer genes (*LAPTM4B* and *YWHAZ*) [43],

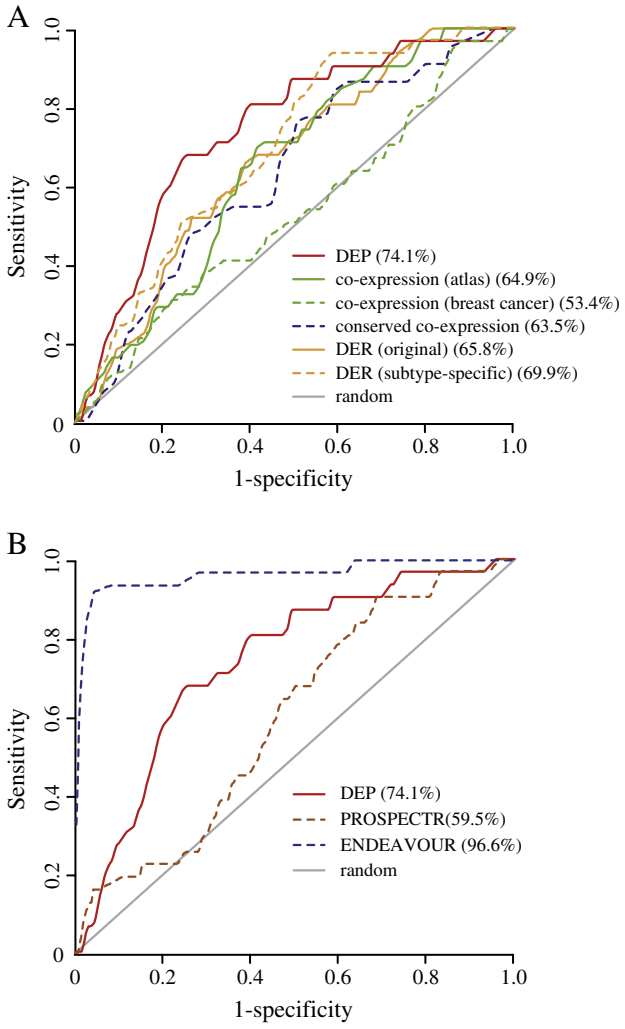

Fig. 4. Comparisons with other prioritization approaches. A) Comparisons with expression-based candidate methods, including co-expression, conserved co-expression, and DER methods. B) Comparisons with PROSPECTR and ENDEAVOUR.

with all 31 known breast cancer genes as the training set. Both genes had higher ranks in our approach (Supplementary Table S5). These results suggested that the DEP-based prioritization approach integrating a large number of expression data sets can be used to identify less well-studied genes.

3.4. Another Case Study: Prostate Cancer

In order to further validate the DEP-based prioritization approach, another case study regarding prostate cancer was performed. A total of 9 prostate cancer-associated case-control expression data sets (Supplementary Table S6) containing 5051 background genes were obtained from the GEO and ArrayExpress databases, and 14 known prostate cancer genes derived from the OMIM database were present in the background genes. As expected, prioritization using known prostate cancer genes significantly outperformed that using random selected genes. There were 78.6% of known prostate cancer genes with relative ranks>0.5 (42.9% were at the interval from 0.8 to 1.0; Supplementary Table S7). In addition, applying the same methods used in our study of breast cancer, we found similar effects of the number of expression data sets, sample size, the number of disease genes, and the number of ANN classifiers as those in the breast cancer study (Supplementary Figs. S3 and S4). Since the platforms used by these expression studies were completely different, the influence of the heterogeneity of platforms was not analyzed. Taken together, successful application of our approach to prostate cancer further supported the efficiency and feasibility of the DEP-based approach for prioritization of candidate genes.

3.5. Application to four breast cancer susceptibility loci

Two recent genome-wide association studies identified four breast cancer susceptibility loci on 14q24.1 (rs999737), 3p (rs4973768), 17q (rs6504950) and 1p11.2 (rs11249433). Using all 31 known breast cancer genes as the positive training set, we applied our approach to prioritize genes located at 10 Mb upstream and 10 Mb downstream of these risk SNPs. The top 10 genes for these four breast cancer susceptibility loci were obtained (Supplementary Table S8). Some genes show strong associations with other cancers. For example, NOTCH2 located in the distal region from the rs11249433 was ranked 7th in the corresponding candidate gene list. It plays an important role in the development and repair of organizations. This gene has been widely demonstrated to be linked with multiple diseases, such as colorectal cancer, chronic lymphocytic leukemia, type 2 diabetes, and pancreatic cancer. THRB ranked 7th in the candidate gene list of rs4973768 is considered as a tumor suppressor involved in cell proliferation, differentiation, and apoptosis. Many studies have found the aberrant methylation or inactivation of THRB gene in renal cell carcinoma, lung cancer, prostate carcinoma, and colorectal cancer. Further, Ling et al. recently demonstrated that THRB shows abnormal expression and frequent hypermethylation of the promoter region in breast cancer [44].

4. Discussion

Transcriptional changes of disease genes caused by the interaction between genetic and environmental factors can result in disease phenotype. Based on the potential relationship between disease genes and transcriptional changes, we developed a novel DEP-based approach which takes advantage of a large number of disease-specific case-control expression data to prioritize candidate genes. Analysis using the leave-one-out cross-validation suggested that the DEP-based prioritization approach can effectively identify candidate disease genes.

In particular, the performance was further improved using subtype-specific expression data, which may be attributed to an evidence that the pathological subtypes of breast cancer are indeed biologically distinct entities [45]. Many studies have demonstrated that breast cancer can be divided into multiple subtypes on the basis of mRNA expression, miRNA expression, and DNA methylation. These subtypes show significant differences in expression patterns, clinical outcomes and survival, which may be caused by different genetic mutations. Therefore, original expression data sets cannot exactly reflect the relationship between disease genes and transcriptional changes. In contrast, subtype-specific expression data sets provide more precise delineation about the relationship. For example, RB1CC1, a key regulator of the tumor suppressor gene RB1, showed the significant improvement of relative rank from 0.24 to 0.79 when subtype-specific expression data sets were used. Despite no direct evidence about the subtype specificity of RB1CC1, its closely associated target RB1 shows strong subtype specificity. Loss of heterozygosity of the RB1 gene has the highest frequency in basal-like tumors but with an obviously low overall frequency, and the significantly different expression patterns of RB1 are present in different subtypes [46]. Consistently, we observed significantly differential expression of RB1CC1 in all basal-like subtype-specific data sets, whereas the differential expression pattern cannot be characterized using original data sets.

More importantly, the performance of our approach showed a continuous improvement with the increase of publicly available expression data sets. Moreover, comparisons with other expression-based prioritization methods further supported the superiority of our approach. The majority of expression-based prioritization methods used only single expression data regardless of a large number of disease-specific expression data. Although the conserved co-expression approach prioritized candidate genes by integrating multiple expression data sets, the performance was obviously dependent on expression data sets selected, that is, integration of more expression data sets could not improve the performance and may even result in limited performance. Through comparison with ENDEAVOUR, we found that this approach based on integration of diverse biological data can largely improve the performance, but it may be biased to the discovery of well-studied disease genes. Conversely, our approach integrating a number of expression data can be used to discover less well-studied genes. This may be attributed to the fact that expression data can be used to infer novel biological hypotheses. Thus, integration of a large number of expression data and various other biological resources may be more effective for prioritization of candidate genes in the future.

Because DEPs were only dependent on whether genes were identified as differentially or non-differentially expressed genes, regardless of up- and down-regulation information as well as the degree of expression changes, we reconstructed DEPs using a binary encoding strategy to consider this information. Nevertheless, the performance was not significantly improved, and the AUC score only increased from 74.1% to 76.0%.

There were some limitations in the DEP-based prioritization approach. First, because the approach was dependent on a large number of expression data, it was not suitable for some diseases with few expression studies. Second, it should be noted that although our approach provides better prioritization capability for less well-studied genes, the probes for these less well-studied genes may not always be included in different microarray platforms. With the development of next-generation sequencing technologies (e.g., RNA-Seq), genome-wide transcriptional levels can be comprehensively detected, independent of the designed probes in microarray. Integrating these expression data sets will be more effective for prioritizing less well-studied genes. Third, although the performance of our approach was robust to the number of known disease genes, it cannot be used for diseases with very few or no known disease genes. Fourth, only expression data were used in our approach. Other biological resources, such as function annotation, protein interaction network, and text information, would be integrated into our approach in future works, which might more efficiently improve the performance of our approach.

In summary, we proposed a novel DEP-based prioritization approach by integrating a large number of disease-specific case-control expression data sets. It will be helpful for identifying disease genes as one of

supplementary prioritization strategies. We believed that integration of dramatically increasing expression data is useful for further improving the results of the prioritization approach.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2011.04.001.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30871394, 30370798 and 30571034), the National High Tech Development Project of China, the 863 Program (Grant No. 2007AA02Z329), the National Basic Research Program of China, the 973 Program (Grant No. 2008CB517302) and the National Science Foundation of Heilongjiang Province (Grant Nos. ZJG0501, 1055HG009, GB03C602-4, HCXB2009019, HCXS2010008 and BMFH060044).

References

[1] S.J. Furney, B. Calvo, P. Larranaga, J.A. Lozano, N. Lopez-Bigas, Prioritization of candidate cancer genes—an aid to oncogenomic studies, Nucleic Acids Res. 36 (2008) e115.
[2] D. Shriner, T.M. Baye, M.A. Padilla, S. Zhang, L.K. Vaughan, A.E. Loraine, Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies, Nucleic Acids Res. 36 (2008) e26.
[3] P. Zhang, J. Zhang, H. Sheng, J.J. Russo, B. Osborne, K. Buetow, Gene functional similarity search tool (GFSST), BMC Bioinformatics 7 (2006) 135.
[4] A. Schlicker, T. Lengauer, M. Albrecht, Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, Bioinformatics 26 (2010) i561–i567.
[5] F.S. Turner, D.R. Clutterbuck, C.A. Semple, POCUS: mining genomic sequence annotation to predict disease genes, Genome Biol. 4 (2003) R75.
[6] J. Freudenberg, P. Propping, A similarity-based method for genome-wide prediction of disease-relevant human genes, Bioinformatics 18 (Suppl 2) (2002) S110–S115.
[7] X. Ma, H. Lee, L. Wang, F. Sun, CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data, Bioinformatics 23 (2007) 215–221.
[8] L. Miozzi, R.M. Piro, F. Rosa, U. Ala, L. Silengo, F. Di Cunto, P. Provero, Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data, PLoS One 3 (2008) e2439.
[9] J.L. Morrison, R. Breitling, D.J. Higham, D.R. Gilbert, GeneRank: using search engine technology for the analysis of microarray experiments, BMC Bioinformatics 6 (2005) 233.
[10] M.A. van Driel, K. Cuelenaere, P.P. Kemmeren, J.A. Leunissen, H.G. Brunner, A new web-based data mining tool for the identification of candidate genes for human genetic disorders, Eur. J. Hum. Genet. 11 (2003) 57–63.
[11] U. Ala, R.M. Piro, E. Grassi, C. Damasco, L. Silengo, M. Oti, P. Provero, F. Di Cunto, Prediction of human disease genes by human-mouse conserved coexpression analysis, PLoS Comput. Biol. 4 (2008) e1000043.
[12] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein-protein interaction network, Bioinformatics 22 (2006) 2800–2805.
[13] J. Chen, B.J. Aronow, A.G. Jegga, Disease candidate gene identification and prioritization using protein interaction networks, BMC Bioinformatics 10 (2009) 73.
[14] K. Lage, E.O. Karlberg, Z.M. Storling, P.I. Olason, A.G. Pedersen, O. Rigina, A.M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau, S. Brunak, A human phenome–interactome network of protein complexes implicated in genetic disorders, Nat. Biotechnol. 25 (2007) 309–316.
[15] M. Oti, B. Snel, M.A. Huynen, H.G. Brunner, Predicting disease genes using protein–protein interactions, J. Med. Genet. 43 (2006) 691–698.
[16] S. Yu, S. Van Vooren, L.C. Tranchevent, B. De Moor, Y. Moreau, Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining, Bioinformatics 24 (2008) i119–i125.
[17] S. Yilmaz, P. Jonveaux, C. Bicep, L. Pierron, M. Smail-Tabbone, M.D. Devignes, Gene-disease relationship discovery based on model-driven data integration and database view definition, Bioinformatics 25 (2009) 230–236.
[18] N. Tiffin, J.F. Kelso, A.R. Powell, H. Pan, V.B. Bajic, W.A. Hide, Integration of text- and data-mining using ontologies successfully selects disease gene candidates, Nucleic Acids Res. 33 (2005) 1544–1552.
[19] R.A. George, J.Y. Liu, L.L. Feng, R.J. Bryson-Richardson, D. Fatkin, M.A. Wouters, Analysis of protein sequence and interaction data for candidate disease gene prediction, Nucleic Acids Res. 34 (2006) e130.
[20] L. Franke, H. van Bakel, L. Fokkens, E.D. de Jong, M. Egmont-Petersen, C. Wijmenga, Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes, Am. J. Hum. Genet. 78 (2006) 1011–1025.
[21] E.A. Adie, R.R. Adams, K.L. Evans, D.J. Porteous, B.S. Pickard, SUSPECTS: enabling fast and effective prioritization of positional candidates, Bioinformatics 22 (2006) 773–774.
[22] J. Chen, E.E. Bardes, B.J. Aronow, A.G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, Nucleic Acids Res. 37 (2009) W305–W311.
[23] C. Perez-Iratxeta, P. Bork, M.A. Andrade-Navarro, Update of the G2D tool for prioritization of gene candidates to inherited diseases, Nucleic Acids Res. 35 (2007) W212–W216.
[24] J.E. Hutz, A.T. Kraja, H.L. McLeod, M.A. Province, CANDID: a flexible method for prioritizing candidate genes for complex human traits, Genet. Epidemiol. 32 (2008) 779–790.
[25] D. Seelow, J.M. Schwarz, M. Schuelke, GeneDistiller—distilling candidate genes from linkage intervals, PLoS One 3 (2008) e3874.
[26] Y. Li, J.C. Patra, Integration of multiple data sources to prioritize candidate genes using discounted rating system, BMC Bioinformatics 11 (Suppl 1) (2010) S20.
[27] E.A. Adie, R.R. Adams, K.L. Evans, D.J. Porteous, B.S. Pickard, Speeding disease gene discovery by sequence based candidate prioritization, BMC Bioinformatics 6 (2005) 55.
[28] S. Kohler, S. Bauer, D. Horn, P.N. Robinson, Walking the interactome for prioritization of candidate disease genes, Am. J. Hum. Genet. 82 (2008) 949–958.
[29] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, Y. Moreau, Gene prioritization through genomic data fusion, Nat. Biotechnol. 24 (2006) 537–544.
[30] J. Chen, E.E. Bardes, B.J. Aronow, A.G. Jegga, ToppGene Suite for gene list enrichment analysis and candidate gene prioritization, Nucleic Acids Res. 37 (2009) W305–W311.
[31] J. Chen, H. Xu, B.J. Aronow, A.G. Jegga, Improved human disease candidate gene prioritization using mouse phenotype, BMC Bioinformatics 8 (2007) 392.
[32] B. Linghu, E.S. Snitkin, Z. Hu, Y. Xia, C. Delisi, Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network, Genome Biol. 10 (2009) R91.
[33] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, J.B. Hogenesch, A gene atlas of the mouse and human protein-encoding transcriptomes, Proc. Natl. Acad. Sci. USA 101 (2004) 6062–6067.
[34] S. Rossi, D. Masotti, C. Nardini, E. Bonora, G. Romeo, E. Macii, L. Benini, S. Volinia, TOM: a web-based integrated approach for identification of candidate disease genes, Nucleic Acids Res. 34 (2006) W285–W292.
[35] R.M. Piro, I. Molineris, U. Ala, P. Provero, F. Di Cunto, Candidate gene prioritization based on spatially mapped gene expression: an application to XLMR, Bioinformatics 26 (2010) i618–i624.
[36] M. Oti, J. van Reeuwijk, M.A. Huynen, H.G. Brunner, Conserved co-expression for candidate disease gene prioritization, BMC Bioinformatics 9 (2008) 208.
[37] R. Chen, A.A. Morgan, J. Dudley, T. Deshpande, L. Li, K. Kodama, A.P. Chiang, A.J. Butte, FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease, Genome Biol. 9 (2008) R170.
[38] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 33 (2005) D514–D517.
[39] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. USA 98 (2001) 5116–5121.
[40] T.A. Thornblad, K.S. Elliott, J. Jowett, P.M. Visscher, Prioritization of positional candidate genes using multiple web-based software tools, Twin Res. Hum. Genet. 10 (2007) 861–870.
[41] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J.S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C.M. Perou, P.E. Lonning, P.O. Brown, A.L. Borresen-Dale, D. Botstein, Repeated observation of breast tumor subtypes in independent gene expression data sets, Proc. Natl. Acad. Sci. USA 100 (2003) 8418–8423.
[42] L. Shi, L.H. Reid, W.D. Jones, R. Shippy, J.A. Warrington, S.C. Baker, P.J. Collins, F. de Longueville, E.S. Kawasaki, K.Y. Lee, Y. Luo, Y.A. Sun, J.C. Willey, R.A. Setterquist, G.M. Fischer, W. Tong, Y.P. Dragan, D.J. Dix, F.W. Frueh, F.M. Goodsaid, D. Herman, R.V. Jensen, C.D. Johnson, E.K. Lobenhofer, R.K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P.K. Wolber, L. Zhang, S. Amur, W. Bao, C.C. Barbacioru, A.B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X.M. Cao, T.A. Cebula, J.J. Chen, J. Cheng, T.M. Chu, E. Chudin, J. Corson, J.C. Corton, L.J. Croner, C. Davies, T.S. Davison, G. Delenstarr, X. Deng, D. Dorris, A.C. Eklund, X.H. Fan, H. Fang, S. Fulmer-Smentek, J.C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P.K. Haje, J. Han, T. Han, H.C. Harbottle, S.C. Harris, E. Hatchwell, C.A. Hauser, S. Hester, H. Hong, P. Hurban, S.A. Jackson, H. Ji, C.R. Knight, W.P. Kuo, J.E. LeClerc, S. Levy, Q.Z. Li, C. Liu, Y. Liu, M.J. Lombardi, Y. Ma, S.R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M.S. Orr, T.W. Osborn, A. Papallo, T.A. Patterson, R.G. Perkins, E.H. Peters, R. Peterson, K.L. Philips, P.S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B.A. Rosenzweig, R.R. Samaha, M. Schena, G.P. Schroth, S. Shchegrova, D.D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K.L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S.J. Walker, S.J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, S. Zhong, Y. Zong, W. Slikker Jr., The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, Nat. Biotechnol. 24 (2006) 1151–1161.
[43] Y. Li, L. Zou, Q. Li, B. Haibe-Kains, R. Tian, C. Desmedt, C. Sotiriou, Z. Szallasi, J.D. Iglehart, A.L. Richardson, Z.C. Wang, Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer, Nat. Med. 16 (2010) 214–218.
[44] Y. Ling, X. Xu, J. Hao, X. Ling, X. Du, X. Liu, X. Zhao, Aberrant methylation of the THRB gene in tissue and plasma of breast cancer patients, Cancer Genet. Cytogenet. 196 (2010) 140–145.
[45] L. Wiechmann, M. Sampson, M. Stempel, L.M. Jacks, S.M. Patil, T. King, M. Morrow, Presenting features of breast cancer differ by molecular subtype, Ann. Surg. Oncol. 16 (2009) 2705–2710.
[46] J.I. Herschkowitz, X. He, C. Fan, C.M. Perou, The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas, Breast Cancer Res. 10 (2008) R75.