A Benchmark of the Marker Gene Tool: MGFM

Khadija El Amrani, Nancy Mah, Andreas Kurtz

1 Berlin-Brandenburg Center for Regenerative Therapies, Charité University Medicine Berlin, 13353, Berlin, Germany

ABSTRACT

Large amounts of microarray experimental data are available in public repositories. Although a variety of tools have been developed to make use of these data, the number of tools that detect marker genes is limited. Identification of marker genes associated with a specific tissue/cell type is a fundamental challenge in genetic and genomic research. In addition to other genes, marker genes are of great importance for understanding the gene function, the molecular mechanisms underlying complex diseases, and may lead to the development of new drug targets. We have previously developed a Bioconductor R package (MGFM) for marker gene detection from microarray data. The tool is freely available from the Bioconductor web site (https://www.bioconductor.org/packages/release/bioc/html/MGFM.html), and it is also provided as an online application integrated into the CellFinder platform (http://cellfinder.org/analysis/marker). In this work, we applied our tool to a public microarray data set from the NCBI’s Gene Expression Omnibus public repository encompassing samples for 12 human tissues. We compared the set of predicted marker genes to a set of tissue-specific genes obtained from the Tissue-specific Gene Expression and Regulation (TiGER) database. Furthermore, we tested the performance of the tool using two different normalization methods, RMA and YuGene. To validate the set of predicted marker genes, tissue-specific genes (gold-standard marker genes) for the examined 12 human tissues were collected from the Tissue-specific Gene Expression and Regulation (TiGER) database. TiGER is a database for generating comprehensive information about human tissue-specific gene regulation, including both expression and regulatory data. The database contains tissue-specific gene expression profiles or expressed sequence tag (EST) data, cis-regulatory module (CRM) data, and combinatorial gene regulation data.

METHODS

Identification of marker genes

To define if a probe set is a potential marker, we sort the corresponding expression values in decreasing order. Next, we define cut-points as those that segregate samples of different types. Each cut-point is defined by the ratio of the expression averages of the groups of samples adjacent to it. The score associated with the first cut-point is defined as a measure of specificity of the probe set. The score values range from 0 to 1. Values near 0 would indicate high specificity and large values closer to 1 would indicate low specificity. For more details about the method, see [1].

Data sources

The microarray expression data set is derived from GEO and it is publicly available with the series number GSE3526 [4]. The data set consists of 36 samples and is derived from 12 human tissues (bone marrow, cervix, heart atrium, kidney cortex, liver, lung, lymph node, midbrain, ovary, prostate, spleen, and testes). Each tissue is represented by three replicate samples. The score details about the method, see [1].

Data normalization

Two methods were used for normalization of the microarray data set, YuGene [5] and the Robust Multiarray Averaging [6] (RMA).
RESULTS

To evaluate the performance of our tool, we collected tissue-specific genes (gold-standard marker genes) for the examined 12 human tissues (bone marrow, cervix, heart atrium, kidney cortex, liver, lung, lymph node, midbrain, ovary, prostate, spleen, and testis) from the TiGER database. For validation of the potential marker sets, only gold-standard marker genes that were also found on the microarray were considered for the validation. This corresponds to a total of 2673 marker genes for the 12 human tissues. Using YuGene, 12354 probe sets out of 54675 (comprising about 22.6 % of all probe sets on the microarray) were identified as potential markers associated with the 12 analyzed tissues. Using RMA, 12580 probe sets out of 54675 (or 23 % of all probe sets on the microarray) were identified as potential markers for all examined tissues. The number of predicted marker probe sets is highly variable between the analyzed tissue types (see Figure 1). The testis shows the largest number of tissue enriched probe sets (n=4213 (RMA), n=4030 (YuGene)), followed by the midbrain (n=2112 (RMA), n=1953 (YuGene)) and the liver (n=1122 (RMA), n=1235 (YuGene)). The predicted marker probe sets for a tissue by both methods RMA and YuGene are mostly the same (see the overlap in Figure 1). Figure 2 shows the number of obtained marker genes from the TiGER database for each tissue as well as the number of correctly identified marker genes using both normalization methods RMA and YuGene. YuGene performed slightly better than RMA. Our tool identified 1026 of the gold-standard marker genes (or 38.4 %) using YuGene, and identified 1013 (or 37.9 %) of the gold-standard marker genes using RMA. The best performance is achieved for testis, where 78 % of the gold-standard marker genes were correctly identified. The lowest precision was obtained for cervix, for which only one marker gene was identified using YuGene.

DISCUSSION

We have previously developed a Bioconductor R package (MGFM) for marker gene detection from microarray data. In this communication, we presented a benchmark of our tool using a public microarray data set, and compared the set of predicted marker genes for the examined 12 tissues to a set of gold-standard marker genes obtained from the TiGER database. The tissue-specific genes in TiGER were selected by comparing expression sequence tags (EST) in 30 human tissues using a binomial test on EST counts. We tested the performance of our tool using two normalization methods, RMA and YuGene. YuGene performed slightly better than RMA. Our tool identified 38.4 % or 37.9 % of the gold-standard marker genes using YuGene or RMA, respectively. The differences between the results of both methods do not appear to be significant. Using both normalization methods, the testis had the largest number of predicted marker probe sets, followed by the midbrain, and the liver, suggesting more specialized molecular functions in these tissues compared to the other tissues. The presented benchmark suffers from some limitations.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Sample IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>bone marrow</td>
<td>GSM80576, GSM80577, GSM80603</td>
</tr>
<tr>
<td>cervix</td>
<td>GSM80633, GSM80634, GSM80635</td>
</tr>
<tr>
<td>heart atrium</td>
<td>GSM80655, GSM80656, GSM80698</td>
</tr>
<tr>
<td>kidney cortex</td>
<td>GSM80686, GSM80687, GSM80689</td>
</tr>
<tr>
<td>liver</td>
<td>GSM80739, GSM80730, GSM80729</td>
</tr>
<tr>
<td>lung</td>
<td>GSM80707, GSM80710, GSM80712</td>
</tr>
<tr>
<td>lymph node</td>
<td>GSM80735, GSM80737, GSM80738</td>
</tr>
<tr>
<td>midbrain</td>
<td>GSM80700, GSM80701, GSM80702</td>
</tr>
<tr>
<td>ovary</td>
<td>GSM80758, GSM80759, GSM80780</td>
</tr>
<tr>
<td>prostate</td>
<td>GSM80805, GSM80806, GSM80824</td>
</tr>
<tr>
<td>spleen</td>
<td>GSM80808, GSM80825, GSM80826</td>
</tr>
<tr>
<td>testis</td>
<td>GSM80853, GSM80868, GSM80869</td>
</tr>
</tbody>
</table>

Table 1: The corresponding samples to the tissues.

First, the gold-standard marker genes from TiGER were calculated based on EST counts. The use of EST counts to quantify gene expression levels is less sensitive than microarray technology. Therefore, the used gold-standard marker gene lists from TiGER are not comprehensive. Second, these transcript-based methods of marker gene determination are not suited for identification of post-translationally modified protein coding genes.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ABBREVIATIONS

GEO: Gene Expression Omnibus
MGFM: Marker Gene Finder in Microarray gene expression data
RMA: Robust Multilayer Average
TiGER: Tissue-specific Gene Expression and Regulation

REFERENCES

**Figure 1:** Number of predicted marker probe sets by MGFM for each of the examined tissues using RMA and YuGene for normalization, and the overlap of markers obtained by both methods for each tissue.

**Figure 2:** Number of obtained gold-standard marker genes from the TiGER database and the number of correctly identified marker genes by our tool for each of the examined tissues, using RMA and YuGene for normalization.