bioRxiv preprint doi: https://doi.org/10.1101/150896. this version posted June 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

GEOracle: Mining perturbation experiments using free text metadata in Gene Expression Omnibus

Djordje Djordjevic [1,2], Yun Xin Chen [1], Shu Lun Shannon Kwan [1], Raymond W. K. Ling [1], Gordon Qian [1], Chelsea Y. Y. Woo [1], Samuel J. Ellis [1] and Joshua W. K. Ho [1,2,*]

[1] Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia
[2] St Vincent's Clinical School, University of New South Wales, Darlinghurst, NSW 2010, Australia

* To whom correspondence should be addressed.

Abstract

Summary: There exists over 1.6 million publicly available gene expression samples across 79,000 data series in NCBI's Gene Expression Omnibus database. Due to the lack of the use of standardised ontology terms to annotate the experimental type and sample type, this database remains difficult to harness computationally without significant manual intervention. In this work, we present an interactive R/Shiny tool called GEOracle that utilises text mining and machine learning techniques to automatically identify perturbation experiments, group treatment and control samples and perform differential expression. We present applications of GEOracle to discover conserved signalling pathway target genes and identify an organ specific gene regulatory network.
Availability: GEOracle is available at http://georacle.victorchang.edu.au/
Contact: jho@victorchang.edu.au
Supplementary information: Supplementary data are available at *BioRXiv*

1 Introduction

The NCBI Gene Expression Omnibus (GEO) is one of the largest public repositories for genome-wide omic data, including mostly transcriptomic data (Barrett *et al.*, 2013). As of March 2017, GEO contains over 79,000 data series (GSE), consisting of over 1.6 million individual gene expression samples (GSM). This database harbours biological insights that are not apparent when studying each data set individually (Rung and Brazma, 2013). Several packages are available to programmatically access GEO data, including GEOquery (Davis and Meltzer, 2007), GEOmetadb (Zhu *et al.*, 2008), compendiumdb (Nandal *et al.*, 2016) and shinyGEO (Dumas *et al.*, 2016), allowing keyword based search and download of GSE and GSM, with few standard analysis options.

One major challenge in effectively reusing public gene expression data is the availability of good quality metadata. The need for standardisation of metadata is the reason for the development of the Minimum Information About a Microarray Experiment (MIAME) standard (Brazma *et al.*, 2001), and more recently the MINSEQE standards for sequencing data (Rung and Brazma, 2013). While some fields in GEO metadata use controlled vocabularies (*e.g.*, species name, gene symbols), the majority of the metadata appears as free text, describing the context of samples (*e.g.*, tissue type or developmental stage) and the experimental design (*e.g.*, perturbation experiment). Although this free text is often readily interpretable by humans, there is no simple means to process this information from GEO in an automated fashion. Ultimately this imposes a major limitation on effectively re-using the huge amount of public data in GEO (Rung and Brazma, 2013). While we believe it is important to push for the use of standard annotations, we nonetheless wish to reuse the large amount of data that exists in GEO.

A gene expression experiment can typically be classified based on its experimental design (*e.g.*, perturbation, time-series and case-control experiments). In many cases, data sets from perturbation experiments (*e.g.*, gene knock-out, signalling stimulation, or physical stimulation) are valuable because they allow us to identify the set of genes that are causally downstream of the perturbation agent. This has important applications in determining signalling pathway targets and regulatory networks (Parikh *et al.*, 2010; Djordjevic *et al.*, 2014; Xiao *et al.*, 2015; Schubert *et al.*, 2016). There are tens of thousands of perturbation studies in GEO, likely containing millions of experimentally determined perturbation data. Nonetheless, currently there is no simple way to determine whether a GSE contains perturbation data. Furthermore, even when a GSE is known to contain perturbation data, it is not trivial to automatically match the treatment samples with their respective control samples since a single GSE may contains multiple treatment and control groups.

In light of these challenges, we use text mining and machine learning techniques to classify GSE that contain perturbation data, and to identify and match the treatment and control samples in a perturbation data set. Text mining of free text metadata has previously been used to identify related experiments through semantic similarity (Galeota and Pelizzola, 2016), and to automatically process large amounts of the GEO database with limited quality control or user oversight (Zinman *et al.*, 2013; Wang *et al.*, 2016). Using our R Shiny tool called GEOracle, we can quickly annotate many perturbation experiments from GEO in a semi-automated fashion with full user control. GEOracle then performs differential expression analysis to identify gene targets of the perturbation agent.

bioRxiv preprint doi: https://doi.org/10.1101/150896. this version posted June 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

2 Implementation

The GEOracle workflow follows the same steps a bioinformatcian would employ when analysing perturbation data on GEO. This begins with identifying whether a GSE is a perturbation experiment. Next comes grouping of replicate samples and identifying the perturbation group relevent for the analysis. Finally the appropriate control group is selected and differential expression analysis is performed. In this section we describe our methodology for performing these steps and evaluate GEOracle's performance on manually curated training and test sets. Given a list of GSE accession numbers, GEOracle begins by extracting their metadata via the R package GEOmetaDB (Zhu *et al.*, 2008).

2.1 Classifying perturbation GSE

To build a classifier for identifying perturbation experiments, we manually curated a training set of 277 randomly selected GSE IDs, which we annotated with the experimental design (Supp. File 1).

Based on 31 manually defined textual features from the free text metadata that can differentiate perturbation experiments (including keywords such as 'knockout', 'KO', 'wildtype', 'WT', 'null', '-/-', 'transgenic', ''TG', Supp. File 3), a support vector machine (SVM) classifier was built to predict perturbation GSE. Performance was maximised by the radial basis function kernel (Fig. S1). When evaluated on our training set by 100 rounds of 10-fold cross-validation with internal feature selection, our model produced a mean Area Under the Receiver Operating Characteristic curve of 0.89, suggesting high sensitivity and specificity.

2.2 Grouping replicate samples

To evaluate our automated grouping of GSM samples and subsequent matching of control and perturbation groups, we manually curated a second set of 73 perturbation GSE (Supp. File 2). Half of these GSE were chosen from the previous training set (including the particularly difficult GSE) and the other half were randomly selected perturbation GSE. We annotated the 832 constituent GSM samples into 259 groups labelled as 'perturbation' or 'control', and paired the 'perturbation' sample groups with their appropriate 'control' groups.

For each identified perturbation GSE, GEOracle groups replicate samples using the available GSM metadata. Replicates could mean biological or technical replicates that together form a unit of analysis for differential expression. GSM titles are processed via a series of string manipulations to remove replicate identifiers and tokenise the titles. A simple hierarchical clustering approach is used, based on Gower distance between tokenised GSM titles, with the tree cut at height 0, resulting in identical GSM titles being assigned to one cluster. The same approach is applied to GSM characteristics to produce a second clustering of samples. Based on these two sample clusterings, we identify the most valid clustering outcome and assign confidences to the output, removing datasets with insufficient metadata or invalid clustering results from further analysis (Fig. S2).

Our multi stage clustering approach produces a grouping sensitivity of 93.2% at the GSE level (meaning every sample in a GSE must be correctly grouped for that GSE to be considered a positive result) based on our training set. All incorrectly clustered GSE can be explained by typographical errors and other anomalies in the metadata. This was an improvement over more naive clustering approaches, based solely on the GSM characteristics, GSM titles, or a simple concatenation of the two, producing sensitivities of 64.4%, 86.3% and 74% respectively (Fig. S3). Although samples can often be grouped by either the titles or the characteristics, the process of deciding which information to use is non-trivial. Fig. S4 shows a complex example where a simple concatenation of GSM titles with GSM characteristics fails to group samples correctly, while our multi-stage decision process succeeds.

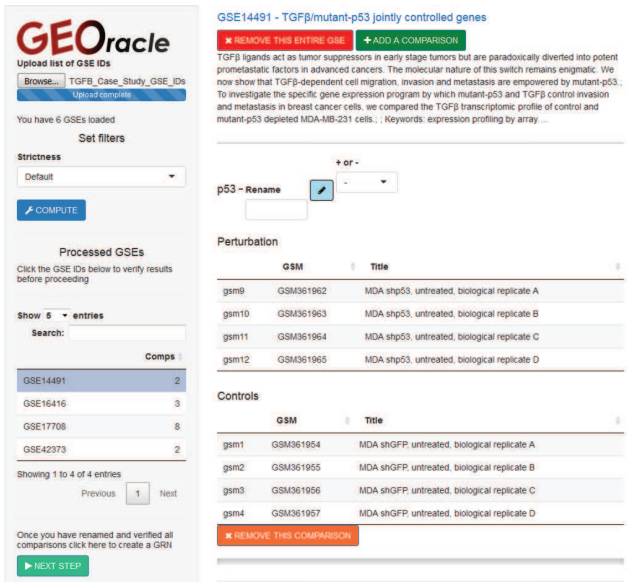

Fig. 1. The GEOracle user interface.

2.3 Classifying sample groups

Both the GSM titles and characteristics were analysed for the presence of 33 textual features that represent molecular concepts that can differentiate 'perturbation' from 'control' samples (Supp. File 4). We trained another SVM classifier to label the groups as 'perturbation' or 'control'. We found the linear kernel for the SVM gave the best results (Fig. S5). We adjust the predicted labels of some groups when only one label is predicted for all samples in a GSE. A confidence associated with the final outcome of group labelling is determined (Fig. S6). We observe a sensitivity of 94.6% for group classification at the GSE level. This is a large improvement over the 73.3% sensitivity produced by the basic approach of choosing the highest scoring label based on the occurence of the subset of 20 features that unambiguously distinguish between 'perturbation' and 'control' samples.

2.4 Matching perturbation with control groups

GEOracle matches each predicted 'perturbation' group to the 'control' group with the lowest Gower distance based on the tokenised GSM titles and characteristics, and determines the confidence for each pairing of groups (Fig. S7). We observe a sensitivity of 83.1% for group matching at the GSE level. Furthermore, we attempt to determine the identity of the perturbation agent and perturbation direction for each group pair by searching for gene names and keywords in the GSM and GSE metadata. The keywords used represent the concepts of addition (i.e. 'overexpress') and removal (i.e. 'knockout') of a perturbation agent. The direction with the most keyword matches becomes the assigned direction.

2.5 Manual adjustment using the graphical user interface

The GEOracle interface (Fig. 1) guides users through the entire process. Importantly the interface allows the user to manually adjust and verify all details of the predicted GSM labels and pairings, and create their own pairings from all GSM within each GSE. This allows the user to be 100% confident in the setup of samples for differential expression analysis.

2.6 Differential expression analyses

The paired 'perturbation' and 'control' groups are then used to compute differential gene expression using GEO2R, which implements the limma pipeline (Ritchie *et al.*, 2015). The results can then be downloaded by the user. GEOracle is currently tailored for microarray data analysis as this

bioRxiv preprint doi: https://doi.org/10.1101/150896. this version posted June 16, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

is the most prevalent data type in GEO, but it can be extended to analyse RNA-seq data or even other functional genomic data sets such as ChIP-seq.

3 Case studies

We used GEOracle to process six GSE containing TGFβ perturbation experiments (Supp. File 5) and discover the consensus target genes of TGFβ signalling stimulation in human cells (Supp. File 6 describes the process in detail). The total time required for classifying the GSE and GSM groups, matching the treatment and control samples, manually verifying the results, downloading the gene expression data from GEO and performing differential expression analysis is less than 12 minutes. This analysis required minimal human intervention and essentially no bioinformatics expertise.

Based on these results we could identify a consensus TGFβ target gene signature in human cells consisting of 82 genes (Fig. S8). Many of the observed trainscriptional changes matched the literature about the TGFβ pathway, including increased transcription of *CTGF*, *JUN*, *JUNB* and *WNT5B*, and repression of *TGFBR3*, *FZD7* and *SPRY1*, (Supp. File 7). A gene ontology analysis of the 82 genes from the consensus signature using g:Profiler (Reimand *et al.*, 2007) showed significant enrichment for the term 'response to transforming growth factor beta', with Benjamini-Hochberg (BH) adjusted p-value of $8.93 \text{ x } 10^{-08}$.

We further used GEOracle to analyse all perturbation microarray data from mouse cardiac tissues. We searched GEO using the following query: "mus musculus"[Organism] AND ("heart"[MeSH Terms] OR heart[All Fields] OR cardiac[All Fields]) AND ("gse"[Filter] AND "Expression profiling by array"[Filter]). This resulted in 851 GSE (Supp. File 8).

Processing these 851 GSE though GEOracle, including manually verifying and modifying the predicted sample comparisons, required approximately 8 hours of user time, again with essentially no bioinformatics expertise required. 164 relevent GSE were included for further processing. The primary reason for exclusion was that the tissue profiled was not of cardiac origin or relevance. We obtained significantly differentially expressed genes for 87 genetic perturbations (i.e. gene knockdown or over expression) and 10 non-genetic factors (diet, chemicals etc.) using standard thresholds (absolute log 2 fold change > 1 and BH adjusted P value < 0.05). GEOracle automatically outputs significant differentially expressed genes as an edge list for gene regulatory network construction. From the genetic perturbation experiments we constructed a gene regulatory network of 23,347 causal and directed relationships between 9,152 genes (Fig. S9). Of these 14,120 were activating relationships and 9,681 were inhibitory. This case study illustrates how we can construct a large organ-specific gene regulatory network from published experimental perturbation data in GEO.

Funding

This work was supported in part by funds from the New South Wales Ministry of Health, a Career Development Fellowship (1105271) from the National Health and Medical Research Council and the National Heart Foundation of Australia, and Amazon Web Services (AWS) Credits for Research.

References

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, **41**(D1), D991–D995.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, **29**(4), 365–371.

Davis, S. and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**(14), 1846–1847.

Djordjevic, D., Yang, A., Zadoorian, A., Rungrugeecharoen, K., and Ho, J. W. K. (2014). How Difficult Is Inference of Mammalian Causal Gene Regulatory Networks? *PLOS ONE*, **9**(11), e111661.

Dumas, J., Gargano, M. A., and Dancik, G. M. (2016). shinyGEO: a web-based application for analyzing gene expression omnibus datasets. *Bioinformatics*, **32**(23), 3679–3681.

Galeota, E. and Pelizzola, M. (2016). Ontology-based annotations and semantic relations in large-scale (epi)genomics data. *Briefings in Bioinformatics*, page doi:10.1093/bib/bbw036.

Nandal, U. K., Kampen, A. H. C. v., and Moerland, P. D. (2016). compendiumdb: an R package for retrieval and storage of functional genomics data. *Bioinformatics*, **32**(18), 2856–2857.

Parikh, J. R., Klinger, B., Xia, Y., Marto, J. A., and Blüthgen, N. (2010). Discovering causal signaling pathways through gene-expression patterns. *Nucleic Acids Research*, **38**(suppl 2), W109–W117.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, **35**(Web Server issue), W193–200.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), e47–e47.

Rung, J. and Brazma, A. (2013). Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, **14**(2), 89–99.

Schubert, M., Klinger, B., Klünemann, M., Garnett, M. J., Blüthgen, N., and Saez-Rodriguez, J. (2016). Perturbation-response genes reveal signaling footprints in cancer gene expression. *bioRxiv*, page doi:10.1101/065672.

Wang, Z., Monteiro, C. D., Jagodnik, K. M., Fernandez, N. F., Gundersen, G. W., Rouillard, A. D., Jenkins, S. L., Feldmann, A. S., Hu, K. S., McDermott, M. G., Duan, Q., Clark, N. R., Jones, M. R., Kou, Y., Goff, T., Woodland, H., Amaral, F. M. R., Szeto, G. L., Fuchs, O., Schüssler-Fiorenza Rose, S. M., Sharma, S., Schwartz, U., Bausela, X. B., Szymkiewicz, M., Maroulis, V., Salykin, A., Barra, C. M., Kruth, C. D., Bongio, N. J., Mathur, V., Todoric, R. D., Rubin, U. E., Malatras, A., Fulp, C. T., Galindo, J. A., Motiejunaite, R., Jüschke, C., Dishuck, P. C., Lahl, K., Jafari, M., Aibar, S., Zaravinos, A., Steenhuizen, L. H., Allison, L. R., Gamallo, P., de Andres Segura, F., Dae Devlin, T., Pérez-García, V., and Ma'ayan, A. (2016). Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nature Communications*, **7**, 12846.

Xiao, Y., Gong, Y., Lv, Y., Lan, Y., Hu, J., Li, F., Xu, J., Bai, J., Deng, Y., Liu, L., Zhang, G., Yu, F., and Li, X. (2015). Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Scientific Reports*, **5**, 10889.

Zhu, Y., Davis, S., Stephens, R., Meltzer, P. S., and Chen, Y. (2008). GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, **24**(23), 2798–2800.

Zinman, G. E., Naiman, S., Kanfi, Y., Cohen, H., and Bar-Joseph, Z. (2013). ExpressionBlast: mining large, unstructured expression databases. *Nature Methods*, **10**(10), 925–926.