Gene set enrichment; a problem of pathways

Matthew N. Davies, Emma L. Meaburn and Leonard C. Schalkwyk

Advance Access publication date 22 September 2010

Abstract

Gene Set Enrichment (GSE) is a computational technique which determines whether *a priori* defined set of genes show statistically significant differential expression between two phenotypes. Currently, the gene sets used for GSE are derived from annotation or pathway databases, which often contain computationally based and unrepresentative data. Here, we propose a novel approach for the generation of comprehensive and biologically derived gene sets, deriving sets through the application of machine learning techniques to gene expression data. These gene sets can be produced for specific tissues, developmental stages or environments. They provide a powerful and functionally meaningful way in which to mine genomewide association and next generation sequencing data in order to identify disease-associated variants and pathways.

Keywords: *gene set enrichment; annotation database; gene expression data; machine learning; next generation sequencing*

INTRODUCTION

Gene set enrichment (GSE) is a computational technique used in the analysis of gene expression data. The technique determines whether a priori defined set of genes show statistically significant differential expression between two sample tissues, time points or conditions [1]. Gene sets are determined by prior biological knowledge relating to co-expression, function, location or known biochemical pathways. The fundamental principle of GSE is that all biochemical pathways are determined by sets of genes and that if that pathway is in any way related to a biological trait then the co-functioning genes should display a higher degree of enrichment compared with the rest of the transcriptome. A focus on the expression of gene sets rather than that of individual genes makes better use of the information generated by a microarray experiment by allowing genes which show only minor differential expression to contribute to the calculation of the enrichment score (ES). The GSE approach should also lead to a greater incidence of replication within array data by identifying the same biological processes underlying a particular phenotype. These arguments can also be applied to the interpretation of multiple weak association signals in genomewide association studies (GWAS) [2, 3].

The most commonly used algorithm to detect the presence of enrichment for a particular gene set is the gene set enrichment analysis (GSEA) technique [1, 4]. GSEA determines whether members of a gene set *S* tend to occur towards the top or the bottom of

Corresponding author. Matthew Davies, Institute of Psychiatry, Box P, De Crespigny Park, London SE5 8AF, UK. Tel: +44 207 848 0969; Fax: +44 207 848 0866; E-mail: matthew.1.davies@kcl.ac.uk

Matthew Davies is a postdoctoral researcher at the Institute of Psychiatry, Kings College London, working in the field of bioinformatics and machine learning. His group has interests in developing technology related to genotyping, allelotyping, methylation, gene expression and bioinformatics.

Emma Meaburn works at the Department of Psychological Sciences at Birkbeck College, University of London as a lecturer in 2010, where her research focus is identifying and understanding the mechanisms by which genetic (DNA sequence variation and gene expression) and epigenetic (DNA methylation) variation accounts for individual differences in behaviour during childhood and adolescence. She is also a visiting Lecturer at the Institute of Psychiatry, King s College London.

Leo Schalkwyk has been at the SGDP Research Centre since the spring of 2000 where he works on functional genomics, identifying genes involved in behaviour in the mouse. His group studies phenotypes related to activity, anxiety and cognition, and have recently undertaken a large genotype-environment interaction study related to depression. They also have interests in developing technology related to genotyping/allelotyping/methylation, gene expression and bioinformatics.

list L, indicating a correlation with a particular phenotype. Calculation of the GSEA requires N, the total number of genes being examined, k, the number of samples, S, the gene set of interest and L, a list containing the N genes ranked by their correlation scores with a specific phenotype ($L = \{g1, g2, \ldots, gN\}$). For each gene set, the P_{hit} and P_{miss} values are calculated. P_{hit} is defined as the difference between the fraction of the genes in S that are present before a given position i while P_{miss} is the fraction of all the N genes (except those in S) that are present before position i across all possible positions i in the list L. The measure of whether there is a significant difference in expression values for a given gene set between two phenotypes is determined by the ES, which is the score of the maximum of $P_{\text{hit}} - P_{\text{miss}}$ over all positions i in the list L. Determining the statistical significance (P-value) of the ES for each gene set requires a permutation test. The two phenotypes are randomly permuted 1000 times, the ES for the gene set is then re-calculated for each permutation and the P-value is estimated as the proportion of the 1000 random permutations that have an ES lower than the ES for the actual experimental data.

Although there are several variations on the original GSEA algorithm (including parametric analysis of gene set enrichment [5] and generally applicable gene set enrichment [6]), all means of calculating enrichment are highly dependent on the nature of the gene sets used. One major determinant on the ES in GSE is simply the size of the gene set. The use of larger sets results in higher statistical power and higher sensitivity where there is only slight enrichment, making them suitable to detect subtle changes in gene expression. Conversely, a large gene set causes the sensitivity to be decreased where there is a greater degree of enrichment. The composition of the gene set is also important as each individual gene will have a varying degree of association with the specified trait that the set is designed to encapsulate [7]. GSE has weak power to detect a differentially expressed gene set where there is a mixture of strongly associated genes and weakly associated genes as the calculated enrichment will not reflect the diversity of the expression values. It is also wrong to assume that genes with large changes in expression values are making a stronger contribution to a pathway than those with smaller changes. Also, some variation in expression levels may simply be a consequence of other signal regulation events (this is arguably a weakness of both the single gene method and of GSE). Here, we assess the current sources of gene sets and how gene expression data may be used to develop methodologies for the creation of new, more specific gene sets for GSE.

CURRENT SOURCES OF GENE SETS

All regions of the genome interact to a greater or lesser extent and it is therefore difficult to represent them as a modular set of pathways that can form the basis of gene sets. In reality, the complex nature of the genome is such that there is no way to establish a precise cut-off point that would determine membership or non-membership of any given set. However, a reasonable criteria by which to define membership must be devised in order for the GSE approach to be implemented. The most common sources of gene sets used to calculate GSE are derived from annotation databases such as GO [8] and KEGG [9]. GO, which uses standardised biological terms to annotate gene products, is the largest of the annotation databases and the one most commonly used for GSE. The database is composed of three main categories: ‘molecular function’, ‘biological process’ and ‘cellular component’ which together comprise approximately 23 000 terms. Each of the three branches can be represented as a directed acyclic graph with each GO term forming a node. Each node in the graph can have several parents (less specific terms) and several children (more specific terms). Annotation of a gene by any node A implies its automatic annotation by all ancestors of A (the set of broader terms related to A by directed paths). Less specific terms are therefore much easier to detect enrichment for using GSE; however, they are typically of less interest to a researcher than a more specific term. All GO annotations are generated with evidence code that records the type of information on which the annotation is based [10]. Annotations are divided into four categories: experimental, computational, indirectly derived from experimental or computational and unknown. Over 95% of these annotations are computationally derived and are associated with pathways through ‘inferred from electronic association’ evidence code [10] (although even with computational annotation, there are still regions of the genome that are not represented within the database). KEGG is a manually curated database, composed of pathway maps representing molecular interaction and reaction networks [9]. Each KEGG pathway defines a set of genes that can be considered for statistical enrichment testing. While KEGG is not as comprehensive as GO, is also not as reliant on inference rather than direct annotation.

GENE SETS USING EXPRESSION DATA

While the use of annotation databases as the basis for the creation of gene sets is entirely valid, there are certain limitations. Gene sets derived in this way can be made specific to a species but not to a particular tissue, developmental stage or array type. It may therefore be possible to further tailor gene sets in a way that accommodate a researcher's point of interest. Another concern is that manual annotation of genes, even with the use of inference, will not be able to keep pace with the genomic variants currently being identified with GWAS and next generation sequencing (NGS) research. An alternative approach would be to use gene expression data as the basis for the creation of new gene sets. Gene sets can be assigned to a particular species but this potentially could be taken further to develop gene sets relevant to a particular tissue (even within sub-regions of the brain, gene expression has been shown to be extremely tissue specific) or developmental stage of the organism. We suggest that the development of gene sets derived from tissue-specific co-expression patterns will be a powerful new way to perform pathway-based analyses of GWAS and NGS data in order to gain insights into biological mechanisms and pathways underlying disease. Such an approach has the potential to uncover many novel disease-associated pathways, as unlike current gene sets, it is not limited by what is already known (or computationally inferred) about gene function. It also provides an immediate link between genetic variation (both coding and non-coding) and function, which is a major limitation in current research. Although using gene expression data will provide a more accurate and comprehensive representation of the genome, there is a risk of circular reinforcement in which genes over expressed within arrays are wrongly included in gene sets and then rediscovered by GSE. The best approach against this would be to use as broad a range of expression data as possible as the basis of the gene sets, an approach which the increasingly large public repository of gene expression data would allow.

MACHINE LEARNING TECHNIQUES

Gene sets can be derived from gene expression data through the use of machine learning techniques. Machine learning has previously been used to cluster genes based on similarities between expression profiles. Various forms of clustering analysis have been used including hierarchical clustering [11] the *k*-means algorithm [12], self organizing maps [13], singular value decomposition [14] and support vector machines [15]. However, the high dimensionality and sheer complexity of microarray data have made it difficult to develop reliable clustering techniques that could be used to determine gene sets [16]. Also, none of these methods provides formal inferences and more importantly may not be representative of gene–gene interactions. Ideally, gene sets developed through clustering techniques would not only utilize all the available information from a microarray but also incorporate any prior biological data that may be relevant to the clustering including genomic and clinical data. One possible way to improve gene clustering and therefore the creation of gene sets would be to develop a network-based approach. A dynamic network can be generated whereby nodes represent individual genes or loci, which can then be interlinked by weighted values. A very simple representation of a gene network can be provided by a Boolean approach, where the network is represented as a binary model in which a gene is either switched on or off, and model the effects of other genes on a specific target gene through a Boolean function [17]. Although this is computationally efficient in comparison with other techniques, the limitations of this approach are that it does not incorporate intermediate levels of gene expression and also that it assumes transitions between gene activation to be simultaneous. Linear additive regulation models revealed certain linear relations in regulatory systems but failed to capture non-linear dynamics aspects of genes regulation [18]. A more advanced approach would be the use of a Neuro Fuzzy Recurrent Network which is based on a combination of neural networks (an efficient modelling technique in machine learning and data mining), fuzzy set theory [19, 20] and the Dynamic Bayesian Network (DBN) which adds dependencies between variables at consecutive time points [21]. The DBN consists of two networks; an initial network containing all the variables and an initial probability distribution and a transition network with a transition probability distribution between time steps. A network-based approach may provide a framework for generating more sophisticated gene sets for the purposes of enrichment. There are several different approaches towards modelling gene networks.

APPLICATIONS OF GSE

GWAS data

One area of research that may benefit from the development of GSE is the area of GWAS. Genomewide association uses highly multiplexed genotyping platforms which systematically scan the genome in many thousands of individuals in order to identify common susceptibility alleles for complex diseases and disorders [22]. In recent years, this approach has been employed on an unprecedented scale and has successfully identified hundreds of disease-associated variants. However, there is disappointment among researchers over the low proportion of known heritability recovered for many traits and disorders. One possible reason for this 'missing heritability' maybe that a proportion of single neucelotide polymorphisms (SNPs) occur within the sub-threshold ($5 \times 10^{-8} < P < 0.05$) tail of statistically significant association and therefore do not survive stringent correction for multiple testing in a typical GWAS analysis. As a consequence, analytical methods that systematically extract and aggregate these multiple weaker association signals from the sum total of GWAS data gathered, and which can then interpret the results in a biologically context, are of enormous interest. An approach already taken by several groups is to treat GWAS data as analogous to gene expression data and perform a pathway-based analysis using the principles of GSEA [2, 3, 23–26].

Of course, challenges exist in the application of GSE to GWAS data, including how to map SNPs to transcripts, how best to account and correct for linkage disequilibrium between SNPs, aggregating measures of association for multiple SNPs in a transcript and dealing with multiple different transcripts of the same gene. Various approaches have already been taken, including the use of all SNPs within a gene as individual entities (GSEA-SNP) [3], assigning the highest statistic value among all SNPs mapped to the gene as the statistic value of the gene [2] or by using SNPs to define a list of significantly associated genes, counting each only once irrespective of the number of significantly associated SNPs that it contains (ALIGATOR) [3]. In spite of the inherent difficulties at interpreting SNP data at a gene level, GSE may still offer the best opportunity to account for the effects of genetic variation on phenotypic traits.

Next generation sequencing

Also relevant to the application of GSE is the use of NGS technology to study genomic variation. The cost of NGS has already come down enough to be comparable to the cost of microarrays 5 years ago, and therefore RNA sequencing has become a feasible method of producing gene expression profiles [27]. RNAseq profiles have compelling advantages over gene expression microarrays. While in microarray methods, fluorescent background severely limits analysis in the lower half of the range of RNA abundance, conversely, RNAseq analysis is based on read counts and therefore its dynamic range is only limited by the depth of the sequencing. Furthermore, although microarrays have become more comprehensive as feature densities have increased, RNAseq is intrinsically more open as an approach. Analyses of splicing, allelic expression and other phenomena such as RNA editing become possible from the same data set, providing extra dimensions to gene expression data that could be incorporated into the creation of gene sets for GSE. The use of these fine grained comparisons will present their own challenges and some work has already been done on accounting for selection bias in GSE analysis of RNAseq data [28]. However, while it is problematic to incorporate these extra dimensions into the annotated approach towards GSE it should be more straightforward to incorporate them into an expression-based approach. Also, as the manual annotation upon which databases like GO and KEGG rely will struggle to keep pace with the expansion of data that NGS technology is now providing, it seems likely that an automated approach to GSE will have increasing application in the future.

CONCLUSION

The processing of expression and sequencing data is essentially a series of data reductions which are necessary to extract meaningful information. In the case of array data, hybridisation information is converted into pixel images, processed to turn those images into probe-level summaries and then summarised further into a matrix of normalised average expression estimates. Following that, the matrix is further filtered to remove extraneous data and differential expression statistics are calculated for each gene. The conversion of individual genes into gene sets may therefore be seen as the last in many stages of data reduction but one that is crucial to determining genuine differential expression between two phenotypes. It is clear the central dogma of GSE, that a single expression value can be assigned to a single gene with a single function, is challenged, firstly by the natural overlap that occurs within biochemical pathways and, secondly, by the genetic variants within the coding region of a gene which can effect expression. Whether these limitations can be addressed is crucial to determining the future utility of GSE. Optimisation of gene sets based on expression data using a network-based approach would considerably benefit enrichment analysis and help to improve our understanding of the biology which underlies complex traits. A major long-term challenge is the integration of genomic data from multiple sources in order to extract maximum value from gene expression or GWAS experiments. These could include multiple sources of annotation, multiple sets of gene expression across species, data on transcription factors and epigenetic data, each bringing some indication of functional relationships with varying specificity and confidence. Only a co-ordinated approach to the analysis will fully exploit the wealth of genomic data that is currently available to us.

Key Points

- GSE is a means of analysing genomic data by studying the effects of groups of genes on a phenotype rather than individual gene expression.
- Currently, gene sets are usually derived from annotation databases such as GO and KEGG. This provides a limited and sometimes inaccurate representation of genomewide gene expression.
- An alternative source of gene sets could be derived from the use of gene expression data through the use of machine learning techniques.
- The use of expression-derived gene sets can allow for analysis to be adjusted to different tissues types, array types or developmental stages.

FUNDING

We are grateful to the funding provided by the Medical Research Council. M.N.D. is a MRC non-clinical research fellow.

References

1. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.
2. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007;**81**(6):1278–83.
3. Holmans P, Green EK, Pahwa JS, *et al*. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* 2009;**85**:13–24.
4. Mootha VK, Lindgren CM, Eriksson KF, *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;**34**:267–73.
5. Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 2005;**6**:144.
6. Luo W, Friedman MS, Shedden K, *et al*. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 2009;**10**:161.
7. Dinu I, Potter JD, Mueller T, *et al*. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* 2007;**8**:242.
8. Barrell D, Dimmer E, Huntley RP, *et al*. The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 2009;**37**:D396–403.
9. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
10. Rhee SY, Wood V, Dolinski K, *et al*. Use and misuse of the gene ontology annotations. *Nat Rev Genet* 2008;**9**: 509–15.
11. Zhang DX, Stromberg AJ, Spiering MJ, *et al*. Coregulated expression of loline alkaloid-biosynthesis genes in Neotyphodium uncinatum cultures. *Fungal Genet Biol* 2009;**46**(8):517–30.
12. de Souto MC, Costa IG, de Araujo DS, *et al*. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* 2008;**9**:497.
13. Katagiri F, Glazebrook J. Pattern discovery in expression profiling data. *Curr Protoc Mol Biol* 2005, Chapter 22:Unit 22 25.
14. Liu Z, Wang M, Alvarez JV, *et al*. Singular value decomposition-based regression identifies activation of endogenous signaling pathways in vivo. *Genome Biol* 2008; **9**:R180.
15. Liu CC, Hu J, Kalakrishnan M, *et al*. Integrative disease classification based on cross-platform microarray data. *BMC Bioinformatics* 2009;**10**(Suppl 1):S25.
16. Clarke R, Ressom HW, Wang A, *et al*. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 2008;**8**: 37–49.
17. Shmulevich I, Dougherty ER, Kim S, *et al*. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* 2002;**18**:261–74.
18. D'Haeseleer P, Wen X, Fuhrman S, *et al*. Linear modeling of mRNA expression levels during CNS development and injury. *Pac Symp Biocomput* 1999:41–52.
19. Maraziotis IA, Dragomir A, Bezerianos A. Gene networks reconstruction and time-series prediction from microarray

data using recurrent neural fuzzy networks. *IET Syst Biol* 2007;**1**:41–50.

20. Bezerianos A, Maraziotis IA. Computational models reconstruct gene regulatory networks. *Mol Biosyst* 2008;**4**: 993–1000.
21. Nam H, Lee K, Lee D. Identification of temporal association rules from time-series microarray data sets. *BMC Bioinformatics* 2009;**10**(Suppl 3):S6.
22. Pare G. Genome-wide association studies–data generation, storage, interpretation, and bioinformatics. *J Cardiovasc Transl Res*;**3**:183–8.
23. Loza MJ, McCall CE, Li L, *et al*. Assembly of inflammation-related genes for pathway-focused genetic analysis. *PLoS One* 2007;**2**:e1035.
24. Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum Genet* 2009;**125**:63–79.
25. Kotti S, Bickeboller H, Clerget-Darpoux F. Strategy for detecting susceptibility genes with weak or no marginal effect. *Hum Hered* 2007;**63**:85–92.
26. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 2008;**92**:265–72.
27. Sultan M, Schulz MH, Richard H, *et al*. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008;**321**:956–60.
28. Young MD, Wakefield MJ, Smyth GK, *et al*. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;**11**:R14.