PLOS ONE

RESEARCH ARTICLE

A novel hypothesis-unbiased method for Gene Ontology enrichment based on transcriptome data

Mario Fruzangohar[1,2]*, Esmaeil Ebrahimie[3,4,5,6], David L. Adelson[1,7]*

1 School of Biological Sciences, The University of Adelaide, Adelaide, South Australia, Australia, **2** School of Agriculture, Food and Wine, The University of Adelaide, Adelaide, South Australia, Australia, **3** Australian Centre for Antimicrobial Resistance Ecology, School of Animal and Veterinary Sciences, The University of Adelaide, Adelaide, South Australia, Australia, **4** School of Medicine, Faculty of Health Sciences, The University of Adelaide, Adelaide, Australia, **5** School of Information Technology and Mathematical Sciences, Division of Information Technology, Engineering and the Environment, University of South Australia, Adelaide, Australia, **6** School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, Australia, **7** Zhendong Australia – China Centre for Molecular Chinese Medicine, The University of Adelaide, Adelaide, South Australia, Australia

* mario.fruzangohar@adelaide.edu.au (MF); david.adelson@adelaide.edu.au (DLA)

Citation: Fruzangohar M, Ebrahimie E, Adelson DL (2017) A novel hypothesis-unbiased method for Gene Ontology enrichment based on transcriptome data. PLoS ONE 12(2): e0170486. doi:10.1371/journal.pone.0170486

Editor: Junwen Wang, Mayo Clinic Arizona, UNITED STATES

Received: August 14, 2016

Accepted: January 5, 2017

Published: February 15, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: MF was supported by PhD divisional scholarship of The University of Adelaide. EE was partially supported by NHMRC APP1061006 in the Alzheimer's Disease Genetics Laboratory, School of Biological Sciences. We acknowledge funding from Nectar to host the web and database server of this study. Nectar is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS). The

Abstract

Gene Ontology (GO) classification of statistically significantly differentially expressed genes is commonly used to interpret transcriptomics data as a part of functional genomic analysis. In this approach, all significantly expressed genes contribute equally to the final GO classification regardless of their actual expression levels. Gene expression levels can significantly affect protein production and hence should be reflected in GO term enrichment. Genes with low expression levels can also participate in GO term enrichment through cumulative effects. In this report, we have introduced a new GO enrichment method that is suitable for multiple samples and time series experiments that uses a statistical outlier test to detect GO categories with special patterns of variation that can potentially identify candidate biological mechanisms. To demonstrate the value of our approach, we have performed two case studies. Whole transcriptome expression profiles of *Salmonella enteritidis* and Alzheimer's disease (AD) were analysed in order to determine GO term enrichment across the entire transcriptome instead of a subset of differentially expressed genes used in traditional GO analysis. Our result highlights the key role of inflammation related functional groups in AD pathology as granulocyte colony-stimulating factor receptor binding, neuromedin U binding, and interleukin were remarkably upregulated in AD brain when all using all of the gene expression data in the transcriptome. Mitochondrial components and the molybdopterin synthase complex were identified as potential key cellular components involved in AD pathology.

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Classifying genes into distinct functional groups through Gene Ontology (GO) is a commonly used and powerful tool for understanding functional genomics and the underlying molecular pathways. The functional genomic changes in bacterial pathogens during disease progression or in emerging highly pathogenic strains are poorly understood.

GO analysis commonly begins with enrichment carried out on a short list of genes with statistically significant differential expression [1–3]. In this method, GO term frequencies in the differentially expressed list of genes are compared to a background control, either GO term frequencies of the whole genome, or another list of genes.

This comparison is usually performed using a one sided Fisher-Exact test or a Hypergeometric distribution. This method is called over-representation analysis (ORA) and is implemented nearly in all current GO analysis tools [3–6].

Using routine GO analysis considers all selected genes contribute equally in the final GO classification. The major limitation to the approach is that the original levels of gene expression can significantly affect protein production and consequently actual GO term enrichment. In addition, genes with low or non-differentially expressed values can participate in final GO enrichment through cumulative effects.

The second limitation of traditional ORA analysis is that it can compare just two samples at a time, but in many situations we need to compare GO enrichments of more than two samples. For example, comparing multiple treatment samples to a control sample, comparing time series of samples from the same tissue and the same species. All of these multi-sample comparisons can help us to better understand causative and conserved biological pathways.

To be able to compare GO enrichments of multiple samples we need a robust statistical framework. To our knowledge, Gene Set Enrichment Analysis (GSEA) [7] is the only well described enrichment method that can be applied to multiple samples. This method and its derivatives has been implemented and tested extensively [7–9]. In this method, expression dataset D with N genes and K samples and their phenotype values/classes are given as input data. Given a set of genes S defined as prior biological knowledge, this method can detect if gene set S is significantly enriched by the input data. In this method, first correlation between each gene and phenotype classes is estimated, then genes are sorted based on the absolute values of the correlations. A cumulative statistic similar to Kolmogorov-Smirnov is calculated for gene set S as the enrichment score. If the enrichment score is higher than a given threshold, gene set S can then be considered as belonging to a significant pathway. The method can be useful for some applications, but it has some limitations in comparative analysis.

The first limitation is that the method depends on a measurable phenotypic value for each sample in order to better estimate correlations and sort the genes. In many applications, the expression profiles of multiple samples have no measurable phenotype, similar to the case studies we described in this study.

Furthermore, GSEA merges expression profiles of K samples into one single sorted gene list. Merging samples eliminate the ability to account for the dynamics and variation pattern of expression profiles across samples. As a result, changing the order of the samples produces the same final sorted gene list. But in many biological studies, especially in time series studies, changing the order of samples can result in different biological interpretations, hence merging is problematic. Availability of Gene Set databases provided by Molecular Signature Database (MSigDB) for all species is another limitation for GSEA analysis. To date, MSigDB only contains gene sets for *Danio rerio*, *Homo sapiens*, *Macac mulatta*, *Mus muculus* and *Rattus norvegicus*.

In order to overcome the limitations of GSEA, we developed an approach to estimate and visualise multiple samples' GO enrichments using their mRNA levels. Our method uses a metric that can identify the most significant biological process(es) or molecular function(s) in a multi sample experiment. We have also developed flexible reports to visualise variation of GO terms across multiple samples.

In this study we show for the first time how mRNA expression levels in bacteria and human can be used to better estimate GO term enrichments. By using mRNA expression levels as coefficients, we are able to consider the impact of low expression level and non-differentially expressed genes such as transcription factors in GO enrichment which are normally discarded in analysis. Furthermore, our approach provides the opportunity to enrich GO terms from the entire transcriptome genome (instead of samples of a short list of genes) and enables us to compare GO enrichments of entire transcriptomes across multiple biological samples.

We implemented the new enrichment method and visual reports on a web server accessible at http://www.comparativego.com. We have used the latest web and database technology (PHP and PostgreSQL) to implement the methods. We are committed to updating the web server database every 12 months. The web server has been tested extensively by different groups from University of Adelaide and worldwide. We recently added support for GO information related to selected eukaryotes including human, zebra fish and yeast.

Bacteria are attractive organisms for GO analysis since they have less post-transcriptional gene silencing compared to animals and plants [10] with mRNA expression levels moderately correlated with protein levels [11]. As the first case study, we applied the new enrichment method to whole transcriptome expression profiling to compare low and high pathogenic strains of one important bacterial pathogen, *Salmonella enteritidis* [12]. The analysis revealed a high level of bacterial-type flagellum-dependent cell motility in the highly pathogenic strain. This mechanism has been well described in *E. coli*, but was not reported in the original work on *S. enteritidis* [12].

As a eukaryotic case study, we employed whole transcriptome GO analysis to profile Alzheimer's Disease (AD) pathology. AD, as the leading cause of dementia, is a major concern worldwide with more than 35 million people affected [13]. There is still no effective treatment available and all therapeutic drugs have failed to show efficacy at the clinical level for individuals with AD symptoms [13]. Whole transcriptome GO analysis that takes into account gene expression levels helped us to develop a novel hypothesis for the molecular mechanisms of AD. We also performed GSEA analysis on this case study dataset and compared its result to our method.

Materials and methods

Incorporation of mRNA expression levels into GO enrichment

Given N *genes* ($g_1 \ldots g_n$) in K samples, we estimate the enrichment score *(ES)* of a GO term *t* in sample *s* $ES_{t,s}$, when expression levels are given as RPKM (Reads per Kilo base per Million Reads)/FPKM (Fragments per Kilo base per Million Reads):

$$ES_{t,s} = \sum_{i=1}^{n} \log_2[e(i,s) + 1] \times I(i,t) \tag{1.1}$$

or as microarray log fold change:

$$ES_{t,s} = \sum_{i=1}^{n} \log_2(2^{e(i,s)} + 1) \tag{1.2}$$

Where *e(i,s)* is the expression level of *gene* g_i in sample *s* and *I(i,t)* is:

$$I(i,t) = \begin{cases} 1, \text{ if } g_i \text{ annotated by GO term } (t) \\ 0, \text{ otherwise} \end{cases}$$

We then define an intermediate value for fold change (*F*) of GO term *t* from sample *s* to sample *s*+1 ($F_{t,s}$):

$$ES_{t,s+1} \Big/ ES_{t,s} \tag{2}$$

Finally, the average fold change of GO term *t* across all samples is defined as:

$$F_t = \sqrt[k-1]{\prod_{s=1}^{k-1} F_{t,s}} \tag{3.1}$$

or log transformed as:

$$F_t = \frac{1}{k-1} \sum_{s=1}^{k-1} \log_2 F_{t,s} \tag{3.2}$$

In general, the most significant GO term associated with an observed expression profile is the one with significantly higher/lower average fold change. It can be identified by an outlier test such as the Grubbs outlier test [14].

GO enrichment is initially estimated at the last (most detailed) level of the GO tree. If there is no significant GO term detected at this level, higher levels (more general levels) of the GO tree are recursively searched until either a significantly represented GO term is found or the highest level of the tree is reached.

Average fold change F_t is sensitive to the order of samples. For example, if we reorder two samples different intermediate fold change values will occur (eq 2). Consequently, the average fold change F_t will change (eq 3).

We also report specific patterns such as GO terms with consistently increasing or decreasing enrichment score between every two consecutive samples:

$$\forall s \in [1..k), \quad ES_{t,s} \leq ES_{t,s+1}$$

GO enrichment proportions versus GO enrichment scores

In sample *s*, the ratio of the enrichment score of GO term *t* to the total of enrichment scores of all GO terms ($t_1 \ldots t_m$) can be considered as the GO enrichment proportion (*EP*) of the GO term:

$$EP_{t,s} = \frac{ES_{t,s}}{\sum_{i=1}^{m} ES_{t_i,s}} \Rightarrow \sum_{i=1}^{m} EP_{t_i,s} = 1$$

GO enrichment proportions are displayed as pie charts on our webserver.

Hypothesis testing tool

Although average fold change and other patterns described in the previous section can detect some patterns in individual GO terms, they cannot tell us whether overall GO term enrichment has significantly changed between two samples. We therefore implemented an integrated

tool on the web server to test the hypothesis of a significant difference between 2 genome/sample GO term enrichment distributions. Specifically, we implemented a Chi-Square test for 2 samples in R [15] and we compared it with the Kolmogorov–Smirnov test [16] for 2 samples. Both tests are non-parametric and are suitable for comparing 2 lists of paired numbers like GO term enrichment scores/proportions between 2 samples.

In order to use these tests, samples were binned based on GO terms (one GO term was treated as one bin), and for each bin, the enrichment score of related GO terms were considered as the count for that bin.

Web application

Methods and algorithms were implemented in our web application [17] using PHP 5 and a PostgreSQL database, running on an Apache webserver in a Linux Fedora environment.

Case study data sets

To demonstrate the biological application of these new methods in global transcriptome GO analysis, expression profiles from two published experiments were used.

The first case study [12] was RNA-Seq global transcriptome data from six strains of *Salmonella enteritidis*, where 3 highly pathogenic strains and 3 low pathogenic strains were compared. The average whole genome expression (RPKM) of 4402 genes of the 3 low pathogenic strains and 3 highly pathogenic strains are presented in S1 File.

For the second case study, whole transcriptome (RNA-Seq) data of AD and normal brains were obtained from Twine et al., 2011 [18]. The RNA was obtained from post-mortem total brains of human normal and AD brains (ID at DNA Data Bank of Japan: SRP004879). The RNA-Seq data was analysed using CLC Genomics workbench (QIAGEN, Finland). Mapping was performed using the following parameter values: mismatch cost: 2, insertion cost: 3, deletion cost: 3, length fraction: 0.8, and similarity fraction: 0.8. RPKM, as expression value, was calculated for 57,773 genes based on the *Homo sapiens* (hg19) reference genome in AD and normal brains. In AD brain samples, 14,720,798 short reads were analysed; 99.98% of the reads were mapped to the reference genome (68.88% to exons and 31.12% to introns). In normal samples, 13,440,858 reads were analysed; 100% of reads were mapped to reference genome (79.27% to exons and 20.73% to introns). Results are presented in S2 File.

To compare the result of our method to GSEA, we used Broad Institute Java application (http://www.broadinstitute.org/gsea). Different parameter sets were tried and following parameters achieved the best result: Permutation Type: Gene Set; Enrichment Statistics; Classic, Metric for Ranking Genes; Ratio of Classes.

Results

Introduction of mRNA expression levels into GO analysis

Combining expression profile data with GO term enrichment provided the opportunity to (a) quantify more accurate GO enrichments, (b) extend analysis coverage from sample-wide to genome-wide, and (c) compare GO enrichments of the same list of genes in multiple biological conditions. By considering the influences of all expressed genes in functional genomics, even those with low levels of expression, we increased the accuracy of GO term analysis.

We have also demonstrated that through the use of an outlier test, we can detect GO terms with extreme variation patterns between samples, indicating possible association with the underlying pathway.

Web application enhancement

Because of the additional computational expense associated with the analysis of the GO distribution of all expressed genes within a genome (global transcriptomics), significant memory and processing resources were required by the Apache web server. To enhance performance and husband system resources we implemented file based caching technology to cache the whole genome GO graphs. When a GO graph is built for the first time, subsequent references to that GO graph, even by other users, are instantaneous. For a better user experience in web applications where long running tasks were performed, we used Ajax technology to implement real time progress bars.

Case studies

Case study 1: Comparison of whole transcriptome based GO enrichment between minimally and highly pathogenic *Salmonella enteritidis*. We used RNA-Seq data for six *Salmonella enteritidis* [12] strains. For each gene in both groups of strains the RPKM counts were averaged.

After submission of both gene lists to the web server, whole transcriptome GO enrichment analysis followed by outlier test was performed (enrichment values and outlier test result is shown in Fig 1). In both Biological Process (BP) and Cellular Component (CC), GOs related to bacterial-type flagellum-dependent cell motility (governed by genes such as *flgB* and *flgC*) were the major differentiating functions between high and low pathogenicity *S. enteritidis*. Flagellated bacteria such as Salmonella and *E. coli* are more mobile and can swim faster. The reversible rotary motor, powered by an ion flux [19], is a significant advantage for bacteria as it provides a tool to rapidly respond to environmental signals and escaping harsh conditions and antibiotics. Interestingly, the GO of "regulation of bacterial-type flagellum-dependent cell motility by regulation of motor speed" (GO ID: 71945, governed by *ycgR* gene) was up-regulated 5.5 fold.

Chemotaxis is another biological process associated with highly pathogenic *S. enteritidis* that was upregulated by more than 5 fold. Genes such as *cheA*, *cheB*, *cheW*, and *cheZ* are central in chemotaxis.

In terms of molecular function, highly pathogenic *S. enteritidis* increase protein-glutamate methylesterase by 8 fold. Protein-glutamate methylesterase is a molecular function in a two-component regulatory system and is regulated by *CheA* and *CheB*. It has been reported that upregulation of protein-glutamate methylesterase and *CheB* significantly contribute in increasing swimming motility and flagella synthesis in *E. coli* [20]. Genetic elements involved in motility are associated with pathogenicity [21]. In addition, it has been demonstrated that these mobile genetic elements (transposons, integrons) increase virulence in animal models as well as colonisation success [21].

Case study 2: Comparison of whole transcriptome based GO enrichment between AD and normal human brain. The results of whole transcriptome GO classification followed by outlier testing in AD and normal brains in Biological Process, Molecular Function and Cellular Component are presented in Fig 2. The most significant functions in AD were inflammation and fatty acid related functions including granulocyte colony-stimulating factor receptor binding, interleukin-1, alcohol dehydrogenase, neuromedin U receptor activity, and norepinephrine transmembrane transporter activity.

Log_2 fold change upregulation of granulocyte colony-stimulating factor receptor binding of AD compared to normal condition is 3.59 (in Molecular Function term). *CSF3* (Colony-stimulating factor-3) is the key member of this functional group. Interleukin-1, Type I receptor

A

Gene Ontology	salmonella_low Enrichment Score	salmonella_high Enrichment Score	Log2 Average Fold Change	Common Genes involved
bacterial-type flagellum-dependent cell motility	5041	1058051	7.713	flgB,flgC,flgE,flgF,flgG,flgH,flgI,flgK,flgL, fliE,fliF,fliG,fliH,fliI,fliJ,fliL,fliM,fliN,fljB, ssaQ,ycgR,
ATP metabolic process	163	8510	5.706	invC,
negative regulation of protein secretion	456	23512	5.688	invE,
negative regulation of bacterial-type flagellum assembly	192	9745	5.665	fliT,
regulation of bacterial-type flagellum-dependent cell motility by regulation of motor speed	55	2540	5.529	ycgR,
regulation of chemotaxis	233	8875	5.251	cheZ,
archaeal or bacterial-type flagellum-dependent cell motility	233	8875	5.251	cheZ,
chemotaxis	4111	145447	5.145	acs,aer,cheA,cheB,cheW,cheZ,fliG,fliJ,fliL,fliM, fliN,ssaQ,tcp,trg,tsr,
bacterial-type flagellum organization	6269	144293	4.525	flgA,flgD,flgG,flhC,fliD,fliJ,fliO,fliP,fliS, fliT,
DNA integration	673	14007	4.379	fliZ,
cysteine metabolic process	217	3557	4.035	sufS,
regulation of protein secretion	370	5405	3.869	sicP,
sulfur compound metabolic process	196	2469	3.655	ynhA,

B

Gene Ontology	salmonella_low Enrichment Score	salmonella_high Enrichment Score	Log2 Average Fold Change	Common Genes involved
protein-glutamate methylesterase activity	56	10075	7.491	cheB,
protein-glutamate O-methyltransferase activity	76	13620	7.486	cheR,
transmembrane signaling receptor activity	1139	57046	5.646	tcp,trg,tsr,
selenocysteine lyase activity	217	3557	4.035	sufS,
oxidoreductase activity, oxidizing metal ions	7200	109071	3.921	dps,
trehalose-phosphatase activity	264	3362	3.671	otsB,
protein tyrosine phosphatase activity	1507	18444	3.613	sptP,wzb,
guanyl-nucleotide exchange factor activity	279	3310	3.568	sopE2,
structural molecule activity	105836	1235950	3.546	eutL,flgI,flgK,flgL,fliE,fljB,iscA,ompA,pduM, pudB,sufA,yadR,
alpha,alpha-trehalose-phosphate synthase (UDP-forming) activity	467	5000	3.42	otsA,
butane-1,4-diamine:2-oxoglutarate aminotransferase activity	102	1013	3.312	oat,
ferroxidase activity	3403	32082	3.237	bfr,
succinylglutamate-semialdehyde dehydrogenase activity	49	430	3.133	astD,
protein transporter activity	9641	82423	3.096	exbB,hofQ,invG,lolB,motA,prgI,ssaC,ssaG,tolQ,

C

Gene Ontology	salmonella_low Enrichment Score	salmonella_high Enrichment Score	Log2 Average Fold Change	Common Genes involved
bacterial-type flagellum filament	2192	963425	8.78	fljB,
bacterial-type flagellum hook	3399	182993	5.751	flgK,flgL,fliD,fliK,
bacterial-type flagellum	520	23423	5.493	cheZ,fliH,fliJ,fliS,
bacterial-type flagellum basal body, MS ring	33	1214	5.201	fliF,
bacterial-type flagellum basal body, distal rod	30	857	4.836	flgG,
bacterial-type flagellum basal body, rod	82	1850	4.496	flgB,flgC,
bacterial-type flagellum basal body, distal rod, L ring	58	922	3.991	flgH,
bacterial-type flagellum basal body, distal rod, P ring	36	539	3.904	flgI,
extracellular space	1502	18442	3.618	sptP,
type III protein secretion system complex	1034	9890	3.258	fliI,invC,ssaN,
extracellular region	167605	1328560	2.987	eno,flgK,fliD,fljB,sipB,sipC,slrP,sopA,sopE2, ssaB,sspH2,yebF,
ATP-binding cassette (ABC) transporter complex, substrate-binding subunit-containing	104	765	2.879	ugpE,

Fig 1. Webserver screenshot from outlier test performed on whole transcriptome of low and high pathogenicity of *Salmonella enteritidis*. (A) Biological Process (B) Molecular Function (C) Cellular Component.

doi:10.1371/journal.pone.0170486.g001

A

Gene Ontology	brain-normal Enrichment Score	brain-AD Enrichment Score		Log2 Average Fold Change	Common Genes Involved
regulation of interferon-gamma production	10.54	252.82	↑	4.584 ✓	CCR7,ISG15,RIPK3,
negative regulation of mitochondrial depolarization	60.1	802.66	↑	3.739 ✓	BCL2,HSH2D,IFI6,NGFR,SRC,
response to type I interferon	29.53	291.29	↑	3.302 ✓	C19orf66,IKBKE,ISG15,MX1,SHMT2,SP100,TRIM56,
regulation of ribonuclease activity	0.28	2.17	↑	2.943 ✓	OAS1,OAS3,
negative regulation of positive chemotaxis	0.07	0.51	↑	2.904 ✓	ANGPT2,
establishment of organ orientation	0.04	0.24	↑	2.681 ✓	IRX4,WNT8A,
canonical Wnt signaling pathway involved in cardiac muscle cell fate commitment	0.04	0.24	↑	2.681 ✓	WNT3A,WNT8A,
negative regulation of proteolysis involved in cellular protein catabolic process	0.02	0.11	↑	2.681 ✓	CDKN2A,
regulation of blood coagulation, intrinsic pathway	0.09	0.56	↑	2.681 ✓	SERPINC1,
regulation of thyroid-stimulating hormone secretion	0.02	0.14	↑	2.681 ✓	PAX8,
dichotomous subdivision of terminal units involved in mammary gland duct morphogenesis	0.02	0.16	↑	2.658 ✓	AREG,AREGB,TFAP2C,
regulation of MyD88-independent toll-like receptor signaling pathway	1.94	11.99	↑	2.627 ✓	IRF7,
epithelium migration	0.03	0.15	↑	2.574 ✓	GRHL2,

B

Gene Ontology	brain-normal Enrichment Score	brain-AD Enrichment Score		Log2 Average Fold Change	Common Genes Involved
granulocyte colony-stimulating factor receptor binding	0.03	0.51	↑	3.874 ✓	CSF3,
anion binding	0.01	0.1	↑	3.781 ✓	TG,
neuromedin U binding	0.1	0.79	↑	2.989 ✓	NMUR2,
interleukin-1, Type I receptor binding	0.02	0.15	↑	2.681 ✓	IL1RN,
interleukin-1 Type I receptor antagonist activity	0.02	0.15	↑	2.681 ✓	IL1RN,
interleukin-1 Type II receptor antagonist activity	0.02	0.15	↑	2.681 ✓	IL1RN,
2'-5'-oligoadenylate synthetase activity	0.38	2.28	↑	2.584 ✓	OAS1,OAS2,OAS3,
gastric inhibitory peptide receptor activity	0.49	2.56	↑	2.385 ✓	GIPR,
thyroid-stimulating hormone receptor activity	0.04	0.17	↑	2.309 ✓	PAX8,TSHR,
acyl-CoA ligase activity	0.03	0.12	↑	2.196 ✓	ACSM1,
immunoglobulin receptor binding	4.76	20.22	↑	2.087 ✓	FES,FGR,IGHD,IGHE,IGHG1,IGHG2,IGHG3,IGHG4, IGHV1OR21-1,IGHV3-23,IGHV3OR16-9,IGHV4OR15-8,IGKC, IGLC1,IGLC2,IGLC3,IGLC6,IGLC7,IGLL1,IGLL5,TRDC,
kynurenine 3-monooxygenase activity	0.14	0.6	↑	2.058 ✓	KMO,
sepiapterin reductase activity	4.34	16.94	↑	1.966 ✓	SPR,
UMP kinase activity	0.81	3.12	↑	1.944 ✓	CMPK2,

C

Gene Ontology	brain-normal Enrichment Score	brain-AD Enrichment Score		Log2 Average Fold Change	Common Genes Involved
granulocyte colony-stimulating factor receptor binding	0.03	0.51	↑	3.874 ✓	CSF3,
anion binding	0.01	0.1	↑	3.781 ✓	TG,
neuromedin U binding	0.1	0.79	↑	2.989 ✓	NMUR2,
interleukin-1, Type I receptor binding	0.02	0.15	↑	2.681 ✓	IL1RN,
interleukin-1 Type I receptor antagonist activity	0.02	0.15	↑	2.681 ✓	IL1RN,
interleukin-1 Type II receptor antagonist activity	0.02	0.15	↑	2.681 ✓	IL1RN,
2'-5'-oligoadenylate synthetase activity	0.38	2.28	↑	2.584 ✓	OAS1,OAS2,OAS3,
gastric inhibitory peptide receptor activity	0.49	2.56	↑	2.385 ✓	GIPR,
thyroid-stimulating hormone receptor activity	0.04	0.17	↑	2.309 ✓	PAX8,TSHR,
acyl-CoA ligase activity	0.03	0.12	↑	2.196 ✓	ACSM1,
immunoglobulin receptor binding	4.76	20.22	↑	2.087 ✓	FES,FGR,IGHD,IGHE,IGHG1,IGHG2,IGHG3,IGHG4, IGHV1OR21-1,IGHV3-23,IGHV3OR16-9,IGHV4OR15-8,IGKC, IGLC1,IGLC2,IGLC3,IGLC6,IGLC7,IGLL1,IGLL5,TRDC,
kynurenine 3-monooxygenase activity	0.14	0.6	↑	2.058 ✓	KMO,
sepiapterin reductase activity	4.34	16.94	↑	1.966 ✓	SPR,
UMP kinase activity	0.81	3.12	↑	1.944 ✓	CMPK2,

Fig 2. Webserver screenshot from outlier test performed on whole transcriptome of normal and Alzheimer's disease of *human* samples. (A) Biological Process (B) Molecular Function (C) Cellular Component.

doi:10.1371/journal.pone.0170486.g002

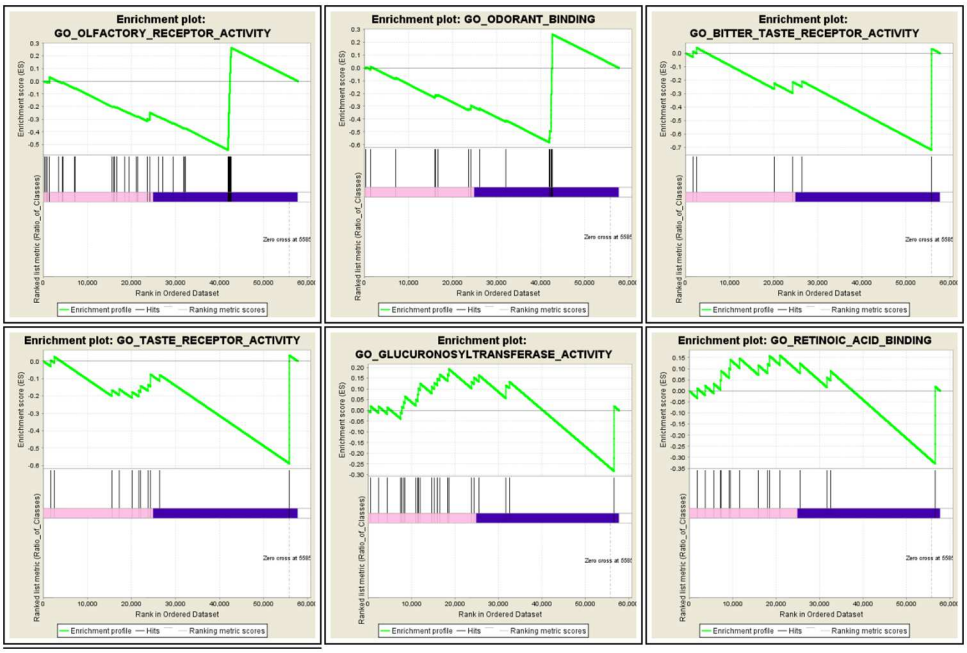

Fig 3. Snapshot of GSEA enrichment result related to molecular function detected in AD.

doi:10.1371/journal.pone.0170486.g003

binding was another inflammation upregulated function which is highly enriched in AD brain (log2 fold change = 2.59, central gene = *IL1RN*).

Cellular components such as the IgA immunoglobulin complex (*IGHA1*, *IGHA2*, and *IGJ*) endoplasmic reticulum membrane (*DHRS7C*, *RHO*), vesicle lumen (*APOB*), and mitochondrial segments (mainly governed by *Bcl3*) were clearly upregulated in AD which highlight the involvement of mitochondria, endoplasmic reticulum, and vesicle formation in AD. Complexes such as Bcl3/NF-kappaB2 complex (governed by *BCL3* and *NFKB2*), vacuolar lumen (governed by *CLN5*), IPAF inflammasome complex (CASP1, CASP4, NLRC4) help to understand the involvement of inflammation and vascular disorder in AD.

GSEA analysis was performed for whole genome of normal and AD samples using tools and parameter set explained in Material and Method. Significant Molecular Functions and Biological Process in AD phenotype are shown in Figs 3 and 4 respectively.

Full result of GSEA is available in S3 File.

Interestingly, molecular functions related to olfactory and sensory receptors stood out in GSEA analysis. There are numerous publications that identified olfactory deficit as early marker of AD at clinical level [22, 23]. At pathological level, several studies [24, 25] have also shown abnormal cellular pattern of the entorhinal cortex and olfactory neurons that disrupt memory function.

Discussion

In this study we showed for the first time how mRNA expression levels in bacteria and human could be used to estimate GO term enrichments. By using mRNA expression levels as coefficients, we were able to include the impact of non-significantly expressed genes in GO enrichment. Furthermore, our approach provided the opportunity to enrich GO terms at the entire transcriptome level (rather than a subset of genes) across multiple biological conditions. The outlier test also detected significant patterns in the data.

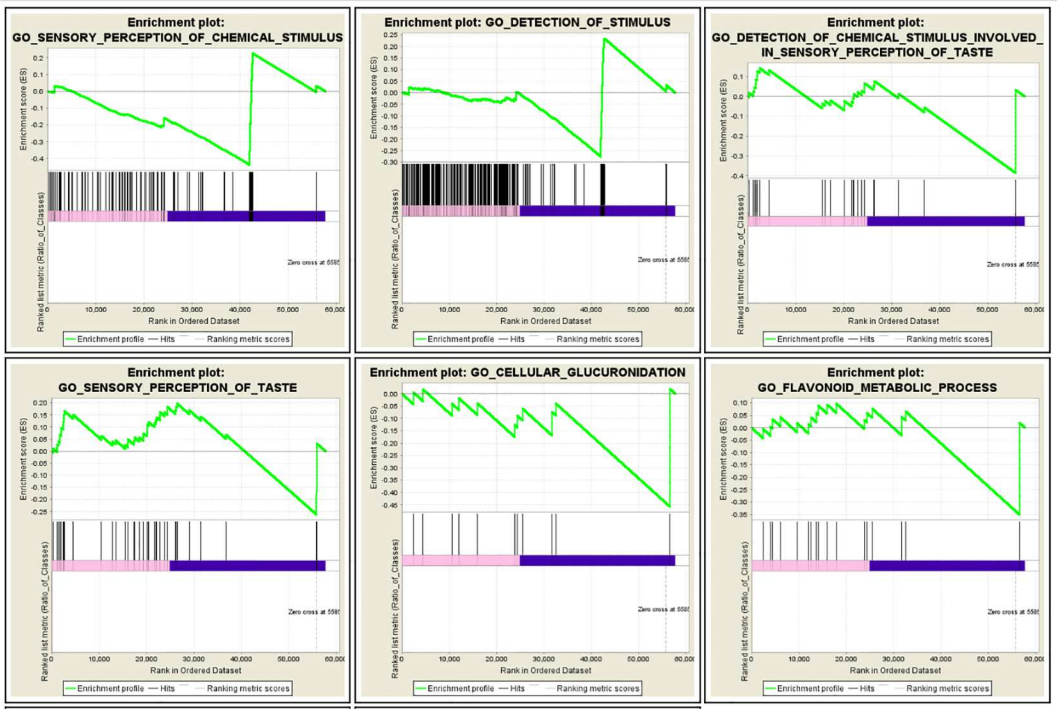

Fig 4. Snapshot of GSEA enrichment result related to biological process detected in AD.

doi:10.1371/journal.pone.0170486.g004

Unlike a previous GSEA method [7], our method is independent of phenotypic data and also reflects the order of samples and is potentially more appropriate for time series experiments such as human degenerative disease or bacterial pathogenesis progress where the condition of a patient/host changes over time. In addition, availability of our method for a much wider range of organisms is another advantage to GSEA that is just available for limited model organisms.

In contrast to other web servers [3, 5], our web server provides interactive visual navigation along the hierarchical structure of GO graphs at all levels of the graph. Furthermore, our web server provides dynamic visual reports (using AJAX technology) including pie charts (to visualize GO enrichment proportions) and bar charts (to visualize over-representation analysis), whereas other web servers present this information in text format or rely on visualization capacity provided by other websites including The European Bioinformatics Institute at http://www.ebi.ac.uk/.

The most significant analytical advantage provided by our web server is the ability to enrich and compare GO terms between multiple gene samples from multiple biological conditions. At present, other web servers [3, 5] can only compare one sample against a control sample. Comparative GO analysis is important as a means to identify underlying biological pathways involved in response to different biological conditions. This is essential if one wishes to identify candidate genes for perturbation experiments.

From a technical point of view, special caching, connection pooling and database query planning and optimization techniques were employed to make the webserver capable of accepting very large lists of genes such as the entire human transcriptome.

We demonstrated the efficiency of our proposed method in prokaryote and eukaryote case studies.

In the case of AD samples, the outlier test revealed upregulation of inflammation related function in AD brain such as granulocyte colony-stimulating factor receptor binding (governed by *CSF3*). *CSF* genes are pro-inflammatory cytokines which are expressed in brain and nervous system disorders and are involved in immunity and inflammation by regulating survival and proliferation. CSF proteins can pass through the blood–brain barrier and influence nervous system activities such as axonal regeneration [26, 27]. *CSF3* can transit between blood and brain and it shows a remarkable increase of its function in AD brain, so we hypothesise that *CSF3* might be tested as a blood based marker of AD in future studies. Another upregulated functional group was Interleukin-1 (IL-1), a pro-inflammatory cytokine that activates many inflammatory processes with important functions in brain neuroimmune responses. IL-1 has been evaluated as a target for therapeutic strategies using Interleukin-1 receptor antagonists in stroke and neural disorders [28–30]. It has also been reported that soluble interleukin-1 receptor increases in the cerebrospinal fluid of AD patients [31]. Interleukin-1 expression in brain can activate caspase-1 and apoptosis. The brain specific mechanisms of action of Interleukin-1 are not yet fully characterised, but may affect glia, endothelia, and neurons [28–30]. Whole transcriptome Gene Ontology based analysis in this study reinforces the hypothesis that inflammatory process are part of the neuropathology in AD. Overall, in comparison to GSEA analysis results that only highlighted already described secondary sensory effects, our method had the ability to introduce new mechanism and new target genes for AD.

Whole transriptome GO comparison of highly pathogenic *Salmonella enteritidis* compared to low pathogenic *S. enteritidis* highlighted the key roles of bacterial-type flagellum-dependent cell motility and chemotaxis in highly pathogenic Salmonella. Chemotaxis is biological process dependent on signal transduction and phosphorylation. We speculate that up regulating GO "Signal transduction by phosphorylation" may allow *Salmonella enteritidis* to more rapidly sense environmental changes and activate more genes through increased phosphorylation activity. It has been documented that chemotaxis is central for virulence and competitive fitness of *Ralstonia solanacearum* [32]. *Ralstonia solanacearum* has a remarkable capability for invading host plant roots from the soil to get amino acids and organic acids [32].

Motility has been identified as key virulence factor in bacteria as many bacteria use flagella to move and cause diseases in humans, animals and plants [33]. In line with our finding on the key roles of flagellum and motility in *S. enteritidis* pathogenicity, it has been demonstrated in *E. coli* that mobile genetic elements and flagellum motility are molecular mechanisms which contribute in increasing *E. coli* pathogenicity to generate a highly adapted pathogen capable of causing a range of diseases in the central nervous system, the gastrointestinal tract, the urinary tract, and blood [34]. Regarding the key roles of flagellum filaments and flagellum-dependent cell motility in bacterial pathogenicity, bacterial genes encoding filament components have been used for vaccine development and therapeutic interventions [35]. *Fli* genes in *S. typhimurium*, *Bacillus subtils* and *E. coli* are the major loci in flagellum biogenesis [36]. Flagellum and motility are also central for invasion of fish hosts by *Vibrio anguillarum* as disruption of the flagellum and loss of motility decreased virulence by 500-fold [37].

We have used our method with mRNA values, but it can also be used with protein values. Using protein abundance would yield even more accurate GO enrichment scores.

The new global transcriptomics, multi-sample GO enrichment methods presented in this report and implemented in the Comparative GO Web application [17] can significantly help to develop new hypotheses for further experiments. The method has the potential to improve bacterial regulatory mechanisms and eukaryotic functional genomics.

Supporting information

S1 File. Whole transcriptome expression levels (RPKM) of low and high pathogenic *Salmonella enteritidis*.
(XLSX)

S2 File. Whole transcriptome expression levels (RPKM) of normal and Alzheimer's disease of human samples.
(XLSX)

S3 File. GSEA analysis result on Alzheimer's disease of human samples.
(XLSX)

Acknowledgments

We would like to thank Brittany Howell and Mohsen Fruzangohar from the University of Adelaide for helping us to revise the manuscript. We acknowledge Nectar to host the web and database server of this study. Nectar is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

Author Contributions

Conceptualization: MF EE DA.

Data curation: MF EE.

Formal analysis: MF EE.

Funding acquisition: EE DA.

Investigation: EE DA.

Methodology: MF.

Project administration: MF DA.

Resources: EE DA.

Software: MF.

Supervision: MF DA.

Validation: MF EE.

Visualization: MF EE DA.

Writing – original draft: MF EE.

Writing – review & editing: MF EE DA.

References

1. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. International journal of plant genomics. 2008; 2008.
2. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005; 21(18):3674–6. doi: 10.1093/bioinformatics/bti610 PMID: 16081474
3. Da Wei Huang BTS, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2008; 4(1):44–57.

4. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nature protocols. 2013; 8(8):1551–66. doi: 10.1038/nprot.2013.092 PMID: 23868073
5. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics. 2004; 20(4):578–80. doi: 10.1093/bioinformatics/btg455 PMID: 14990455
6. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research. 2010:gkq973.
7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(43):15545–50. doi: 10.1073/pnas.0506580102 PMID: 16199517
8. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail—advanced gene set enrichment analysis. Nucleic acids research. 2007; 35(suppl 2):W186–W92.
9. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics. 2007; 23(23):3251–3. doi: 10.1093/bioinformatics/btm369 PMID: 17644558
10. Cogoni C, Macino G. Post-transcriptional gene silencing across kingdoms. Current opinion in genetics & development. 2000; 10(6):638–43.
11. Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010; 329(5991):533–8. doi: 10.1126/science.1188308 PMID: 20671182
12. Shah DH. RNA-Seq reveals differences in the global transcriptome between high-and low-pathogenic Salmonella Enteritidis strains. Applied and environmental microbiology. 2013:AEM. 02740–13.
13. Laske C, Stellos K, Kempter I, Stransky E, Maetzler W, Fleming I, et al. Increased cerebrospinal fluid calpain activity and microparticle levels in Alzheimer's disease. Alzheimer's & Dementia. 2015; 11 (5):465–74. http://dx.doi.org/10.1016/j.jalz.2014.06.003.
14. Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics. 1969; 11(1):1–21.
15. Team RC. R: A language and environment for statistical computing. R foundation for Statistical Computing. 2005.
16. Smirnov N. Table for estimating the goodness of fit of empirical distributions. The annals of mathematical statistics. 1948:279–81.
17. Fruzangohar M, Ebrahimie E, Ogunniyi AD, Mahdi LK, Paton JC, Adelson DL. Comparative GO: A Web Application for Comparative Gene Ontology and Gene Ontology-Based Gene Selection in Bacteria. PloS one. 2013; 8(3):e58759. doi: 10.1371/journal.pone.0058759 PMID: 23536820
18. Twine NA, Janitz K, Wilkins MR, Janitz M. Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease. PLoS ONE. 2011; 6(1): e16266. doi: 10.1371/journal.pone.0016266 PMID: 21283692
19. Berg HC. The Rotary Motor of Bacterial Flagella. Annual Review of Biochemistry. 2003; 72(1):19–54.
20. Ling H, Kang A, Tan MH, Qi X, Chang MW. The absence of the luxS gene increases swimming motility and flagella synthesis in Escherichia coli K12. Biochemical and Biophysical Research Communications. 2010; 401(4):521–6. http://dx.doi.org/10.1016/j.bbrc.2010.09.080. PMID: 20875395
21. Hacker J, Blum-Oehler G, Mühldorfer I, Tschäpe H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Molecular microbiology. 1997; 23(6):1089–97. PMID: 9106201
22. Devanand D, Michaels-Marston KS, Liu X, Pelton GH, Padilla M, Marder K, et al. Olfactory deficits in patients with mild cognitive impairment predict Alzheimer's disease at follow-up. American Journal of Psychiatry. 2000; 157(9):1399–405. doi: 10.1176/appi.ajp.157.9.1399 PMID: 10964854
23. Doty RL, Reyes PF, Gregor T. Presence of both odor identification and detection deficits in Alzheimer's disease. Brain research bulletin. 1987; 18(5):597–600. PMID: 3607528
24. Hyman BT, Van Hoesen GW, Damasio AR, Barnes CL. Alzheimer's disease: cell-specific pathology isolates the hippocampal formation. Science. 1984; 225(4667):1168–70. PMID: 6474172
25. Talamo BR, Rudel R, Kosik KS, Lee VM-Y, Neff S, Adelman L, et al. Pathological changes in olfactory neurons in patients with Alzheimer's disease. 1989.

26. Chitu V, Stanley ER. Colony-stimulating factor-1 in immunity and inflammation. Current Opinion in Immunology. 2006; 18(1):39–48. http://dx.doi.org/10.1016/j.coi.2005.11.006. doi: 10.1016/j.coi.2005.11.006 PMID: 16337366

27. Franzen R, Bouhy D, Schoenen J. Nervous system injury: focus on the inflammatory cytokine 'granulocyte-macrophage colony stimulating factor'. Neuroscience Letters. 2004; 361(1–3):76–8. http://dx.doi.org/10.1016/j.neulet.2003.12.018. PMID: 15135897

28. Allan SM, Tyrrell PJ, Rothwell NJ. Interleukin-1 and neuronal injury. Nat Rev Immunol. 2005; 5(8):629–40. doi: 10.1038/nri1664 PMID: 16034365

29. Rothwell N. Interleukin-1 and neuronal injury: mechanisms, modification, and therapeutic potential. Brain, Behavior, and Immunity. 2003; 17(3):152–7. http://dx.doi.org/10.1016/S0889-1591(02)00098-3. PMID: 12706413

30. Relton JK, Rothwell NJ. Interleukin-1 receptor antagonist inhibits ischaemic and excitotoxic neuronal damage in the rat. Brain Research Bulletin. 1992; 29(2):243–6. http://dx.doi.org/10.1016/0361-9230(92)90033-T. PMID: 1388088

31. Garlind A, Brauner A, Höjeberg B, Basun H, Schultzberg M. Soluble interleukin-1 receptor type II levels are elevated in cerebrospinal fluid in Alzheimer's disease patients. Brain Research. 1999; 826(1):112–6. http://dx.doi.org/10.1016/S0006-8993(99)01092-6. PMID: 10216202

32. Yao J, Allen C. Chemotaxis Is Required for Virulence and Competitive Fitness of the Bacterial Wilt Pathogen Ralstonia solanacearum. Journal of Bacteriology. 2006; 188(10):3697–708. doi: 10.1128/JB.188.10.3697-3708.2006 PMID: 16672623

33. Josenhans C, Suerbaum S. The role of motility as a virulence factor in bacteria. International Journal of Medical Microbiology. 2002; 291(8):605–14. http://dx.doi.org/10.1078/1438-4221-00173. PMID: 12008914

34. Croxen MA, Finlay BB. Molecular mechanisms of Escherichia coli pathogenicity. Nat Rev Micro. 2010; 8(1):26–38. http://www.nature.com/nrmicro/journal/v8/n1/suppinfo/nrmicro2265_S1.html.

35. Craig L, Pique ME, Tainer JA. Type IV pilus structure and bacterial pathogenicity. Nat Rev Micro. 2004; 2(5):363–78. http://www.nature.com/nrmicro/journal/v2/n5/suppinfo/nrmicro885_S1.html.

36. Gijsegem F, Gough C, Zischek C, Niqueux E, Arlat M, Genin S, et al. The hrp gene locus of Pseudomonas solanacearum, which controls the production of a type III secretion system, encodes eight proteins related to components of the bacterial flagellar biogenesis complex. Molecular microbiology. 1995; 15(6):1095–114. PMID: 7623665

37. O'Toole R, Milton DL, Wolf-Watz H. Chemotactic motility is required for invasion of the host by the fish pathogen Vibrio anguillarum. Molecular microbiology. 1996; 19(3):625–37. PMID: 8830252