G3 Genes | Genomes | Genetics

SOFTWARE AND DATA RESOURCES

CGSEA: A Flexible Tool for Evaluating the Associations of Chemicals with Complex Diseases

Shiqiang Cheng, Mei Ma, Lu Zhang, Li Liu, Bolun Cheng, Xin Qi, Chujun Liang, Ping Li, Om Prakash Kafle, Yan Wen, and Feng Zhang[1]

Key Laboratory of Trace Elements and Endemic Diseases of Ministry of Health, School of Public Health, Health Science Center, Xi'an Jiaotong University, Xi'an, P. R. China

ORCID ID: 0000-0001-8427-0312 (S.C.)

ABSTRACT The etiology of many human complex diseases or traits involves interactions between chemicals and genes that regulate important physiological processes. It has been well documented that chemicals can contribute to disease development through affecting gene expression *in vivo*. In this study, we developed a flexible tool CGSEA for scanning the candidate chemicals associated with complex diseases or traits. CGSEA only need genome-wide summary level data, such as transcriptome-wide association studies (TWAS) and mRNA expression profiles. CGSEA was applied to the GWAS summaries of attention deficiency/hyperactive disorder, (ADHD), autism spectrum disorder (ASD) and cervical cancer. CGSEA identified several significant chemicals, which have been demonstrated to be involved in the development or treatment of ADHD, ASD and cervical cancer. The CGSEA program and user manual are available at https://github.com/ChengSQXJTU/CGSEA.

KEYWORDS
complex diseases
chemicals
genome-wide association study
gene set enrichment analysis

The pathogenesis of many human complex diseases or traits arise from interactions between environmental factors and genes that regulate important physiological processes (Olden and Wilson 2000). Chemicals in the environment play critical roles in the etiology of many human complex diseases. For example, benzene is a ubiquitous chemical in our living environment. It can cause acute leukemia and other hematological cancers (Smith 2010). Arsenic contributes to the development of diabetes (Navasacien *et al.* 2005). However, the traditional methods used to explore the interactions between chemicals and complex diseases have some limitations, such as elucidating the molecular mechanisms of action of environmental chemicals, developing methods to predict toxicity effectively and understanding the genetic basis of differential susceptibility (Mattingly *et al.* 2004). In addition, environmental exposure of chemicals is usually mixed. Therefore, it is difficult to accurately measure exposure levels *in vivo*.

It has been well documented that chemicals generate biological effects through affecting gene expression *in vivo*. For example, previous study have observed gene expression changes that were associated with occupational benzene exposure in the peripheral blood mononuclear cell, such as *CXCL16*, *ZNF331*, *JUN* and *PF4* (McHale *et al.* 2009). A previous study suggested that the expression of dispersed genes may be prone to environmental stimuli while that of clustered genes may be resistant and concluded that environmental components were able to account for most of the positional variation in gene expression changes (Choi and Kim 2007). Comparative Toxicogenomics Database (CTD) is a well-known database, which includes extensive annotations of associations between chemicals and gene expression. CTD was established based on the published high-throughput experimental data and curate toxicologically important genes. By searching references with multiple, large vocabularies, the contributors are compiling a more comprehensive literature set that is relevant to the effects of chemicals on gene expression (Mattingly *et al.* 2006).

Genome-wide association study (GWAS) methodology has advanced such that it is now a powerful tool for the dissection of more complex genetic architectures of human diseases or traits (Mccarthy *et al.* 2008). It is well known that gene expression is under genetic control and a large part of the candidate loci identified by GWAS affect diseases by regulating gene expression (Dimas *et al.* 2009). This motivates the development of transcriptome-wide association studies

doi: https://doi.org/10.1534/g3.119.400945
Manuscript received November 25, 2019; accepted for publication January 10, 2020; published Early Online January 14, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.11604990.
[1]Corresponding author: Key Laboratory of Trace Elements and Endemic Diseases of Ministry of Health, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, P. R. China 710061. E-mail: fzhxjtu@xjtu.edu.cn.

(TWAS), which is a promising approach to evaluate the expression associations of gene with complex diseases or traits by only using GWAS summaries and take tissue specificities into consideration (Gusev *et al.* 2016). For instance, Gusev *et al.* used TWAS and identified 69 new genes associated with obesity-related traits in blood and adipose tissue (Gusev *et al.* 2016). In this study, based on the chemical-gene interaction networks, we developed a flexible tool CGSEA, which is capable to detect the associations between chemicals and complex diseases or traits utilizing high-throughput omics summary statistics (such as TWAS and mRNA expression profiles). We applied CGSEA to publicly available GWAS summaries of attention deficiency/hyperactive disorder (ADHD), autism spectrum disorder (ASD) and cervical cancer to illustrate the good performance of CGSEA.

METHODS

Implementation

Step1 - Chemical-gene expression annotation dataset: The relationships between chemicals and gene expression changes were obtained from the CTD (http://ctdbase.org/downloads/), including organic chemicals, polycyclic compounds, biological factors and enzymes and coenzymes. Currently, CTD provides over 1,929,106 interactions between 13,151 chemicals and 48,092 genes in 591 organisms. Specific for this study, a total of 1,788,149 annotation terms of chemical-gene pairs driven from human and mice were used in this study. We finally generated 11,190 chemicals related gene sets. The overview of the information retrieval process of CTD can be found in the previous study (Mattingly *et al.* 2006).

Step2 –Gene expression association testing statistics of complex diseases: In this study, we used the TWAS expression association testing statistics (TWAS *Z-score*) calculated by the FUSION software (http://gusevlab.org/projects/fusion/) (Gusev *et al.* 2016). First, TWAS was conducted to test the associations between target diseases and the gene expression levels imputed by the prediction models of FUSION (Gusev *et al.* 2016). Briefly, for a given gene, the SNP-expression weights in the 1-Mb cis loci of the gene were first computed with the Bayesian sparse linear mixed model (BSLMM) (Zhou *et al.* 2013). The imputed gene expression data can be viewed as a linear model of genotypes with weights based on the correlation between SNPs and gene expression in the training data while accounting for linkage disequilibrium (LD) among SNPs (Gusev *et al.* 2016). The gene expression weights were then combined with summary-level GWAS results to calculate the association statistics between gene expression levels and each of the disease. Specific for this study, the expression weights of brain RNA-seq and whole blood RNA array were downloaded (http://gusevlab.org/projects/fusion/), and used as reference data in the TWAS of ASD and ADHD. The expression weights of cervical squamous cell carcinoma RNA-seq from The Cancer Genome Atlas (TCGA) and whole blood RNA array were downloaded (http://gusevlab.org/projects/fusion/), and used as reference data in the TWAS of cervical cancer. The expression weights reference data of brain RNA-seq, whole blood RNA array and cervical squamous cell carcinoma RNA-seq contain 5419, 4700 and 1117 genes respectively. Let $L_i$denote the TWAS statistic (Z-score) of the *i*th gene. All genes are ranked by sorting $L_{zi}$from maximum to minimum ($L_1^s \geq L_2^s \geq ....L_n^s$), denoted as$L^s = [L_1^s, L_2^s, ...L_n^s]$. Additionally, the gene expression association testing statistics can also be driven from gene expression profile studies.

Step3 – Chemical related gene set enrichment analysis: For a given chemical related gene set C with N_C genes, let $g_i$denotes the i*th* gene of the gene set C. Let ES^C denote the enrichment scores (ES) of gene set C, which was calculated by weighted Kolmogorov–Smirnov-like running sum statistic in gene set enrichment analysis (GSEA) (Subramanian *et al.* 2005, Wang *et al.* 2007), defined by

$$ES^C = \max_{1 \leq j \leq N} \left\{ \sum_{g_i \in C, i \leq j} \frac{|L_i^s|^w}{N_R} - \sum_{g_i \notin C, i \leq j} \frac{1}{N - N_C} \right\},$$

where $N_R = \sum_{g_i \in C} |L_i^s|^w$. N denotes the total number of genes. w is a parameter giving higher weights to genes with extreme statistics. j denotes the gene set size (number of genes) related to a given chemical. CGSEA calculates the ES^C by walking down the ranked list L of genes, increasing a running-sum statistic when a gene is in the gene set C and decreasing it when it is not. Without loss of generality, w was assigned to be 1 in this study. For statistic tests, permutations were conducted to obtain the null distribution of ES^C(denoted as ES^C_{null}), through randomly shuffling the gene labels. Let ES^C_{tnull} denote the ES value of gene set C of *t*th permutation. After P times permutations, we obtained the null distribution ofES^C_{null}, denoted as $ES^C_{null} = [ES^C_{1null}, ES^C_{2null}, ... ES^C_{pnull}]$. The observed ES ($ES^C$) of the gene set C was normalized by the mean value and standard deviation of permutated ES (ES^C_{null}), defined by

$$NES^C = \frac{ES^C - mean(ES^C_{null})}{SD(ES^C_{null})}$$

Where NES^Cdenoted the normalized ES of the gene set C. Let NES^C_{null} denoted the null distribution of NES^C. NES^C_{null} was defined as $NES^C_{null} = [NES^C_{1null}, NES^C_{2null}, ... NES^C_{pnull}]$, which could be calculated from P permutations using the similar formula:

$$NES^C_{inull} = \frac{ES^C_{inull} - mean(ES^C_{null})}{SD(ES^C_{null})}$$

For the given chemical related gene set C, the empirical P were calculated from the observed NES^C and NES^C_{null} following the widely used approach (Wang *et al.* 2007, Wen *et al.* 2016). We developed a tool CGSEA to implement the approach proposed by this study. In GSEA, the gene sets are defined based on prior biological knowledge, such as published information about biochemical pathways or coexpression in previous experiments (Subramanian *et al.* 2005). In addition, the traditional GSEA usually obtained the gene expression statistics from mRNA expression profiles. In CGSEA, a gene set is any group of genes that share a particular chemical, and the aim is to determine whether that chemical has a role in the phenotype of interest. Meanwhile, the gene expression statistics were computed by TWAS, which is not susceptible to the environmental confounders that may influence expression. However, despite those differences, the underlying statistical structure (weighted Kolmogorov–Smirnov-like statistic) is essentially the same. To facilitate the application of CGSEA, the CTD chemical-gene annotation file used by this study has been included in the CGSEA package. Figure 1 presents the general analytical procedures of CGSEA.

Application to ADHD, ASD and cervical cancer

The GWAS summaries of ADHD (19 099 cases and 34 194 controls) and ASD (7 387 cases and 8 567 controls) were downloaded from the

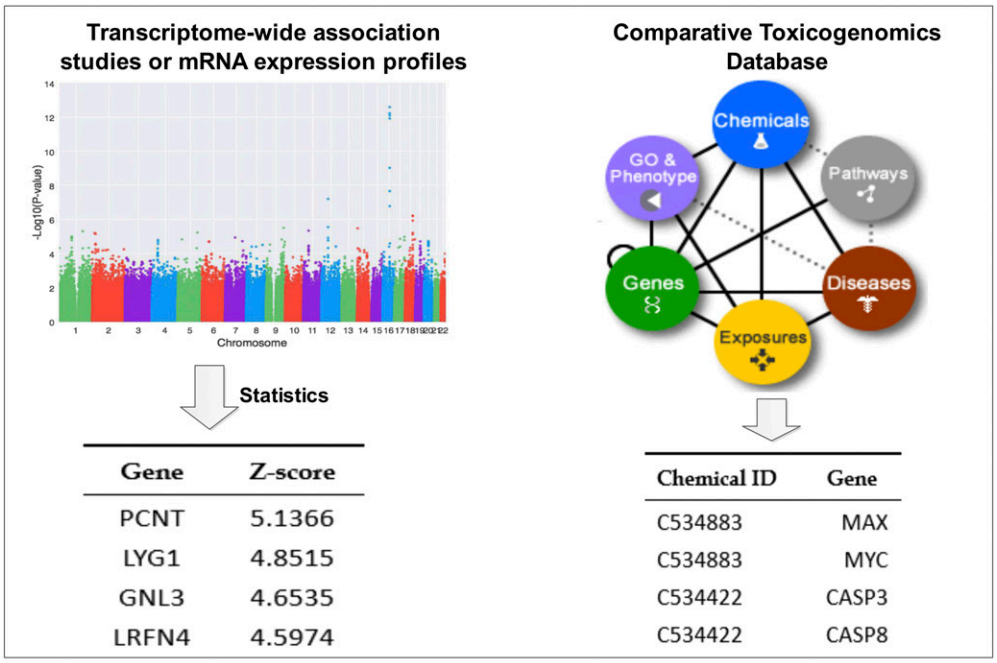

Figure 1 The general analytical procedures of CGSEA.

Table 1 List of top three chemicals identified by CGSEA for ADHD, ASD and cervical cancer

Disorders	Chemicals	Empirical *P* value
ADHD	Crizotinib	0.0002
	Ketoconazole	0.0004
	Methylazoxymethanol Acetate	0.0006
ASD	Ptaquiloside	0.0008
	Ro 31-8220	0.0010
	Ethoxyquin	0.0024
Cervical cancer	Uranium	0.0042
	4-toluidine	0.0068
	Vitamin E	0.0130

Note: attention deficiency/hyperactive disorder, ADHD; autism spectrum disorder, ASD.

Psychiatric GWAS Consortium (PGC) website (https://www.med.unc.edu/pgc/ results-and-downloads). The GWAS summaries of carcinoma *in situ* of cervix uteri were derived from the UK Biobank, including 1 992 patients and 243 502 healthy controls of European ancestry. TWAS was conducted by the FUSION software (http://gusevlab.org/projects/fusion/) (Gusev *et al.* 2016). For the TWAS of ADHD and ASD, the gene expression weight references of whole blood and brain tissues were used. For the TWAS of cervical cancer, the gene expression reference weight of cervical squamous cell carcinoma was used. 5,000 permutations were conducted by CGSEA in this study.

Data availability

All data used in this manuscript are freely available and published online. The GWAS summaries of ADHD and ASD can be found on the Psychiatric GWAS Consortium (PGC) website (https://www.med.unc.edu/pgc/ results-and-downloads). The GWAS summaries of carcinoma *in situ* of cervix uteri can be found on the UK Biobank under the accession D06 (http://geneatlas.roslin.ed.ac.uk/downloads/). Supplemental material available at figshare: https://doi.org/10.25387/g3.11604990.

RESULTS AND DISCUSSION

Table 1 summarizes the top three significant chemicals identified by CGSEA for ADHD, ASD and cervical cancer, respectively. The functional relevance of some identified chemicals with ADHD, ASD and cervical cancer have been reported by previous study. For example, prenatal exposure to methylazoxymethanol acetate (P = 0.0006) lead to alterations in the medial prefrontal cortex indicative of a compromise in information processing (Goto and Grace 2006). Ro-31-8220 (P = 0.0010) is one of AKT inhibitors. Chen *et al.* have suggested that IGF-I/PI3K/AKT/mTOR pathway has potency in the diagnosis and treatment of ASD (Chen *et al.* 2014). Antioxidant vitamin (vitamins A, C, and E (P = 0.0130)) intake was suggested to decrease the risk of cervical cancer (Kim *et al.* 2010). The increase in the incidence of precancerous lesions of the cervix in areas near the borders with the former Yugoslavia during 1997-1999 may be influenced by environmental factors such as exposure to depleted uranium (P = 0.0042) (Papathanasiou *et al.* 2005).

In this study, we developed a flexible tool CGSEA for scanning the candidate chemicals associated with complex diseases or traits. CGSEA has two advantages. First, our approach only need genome-wide summary level data (such as the summaries of TWAS and mRNA expression profiles), which are usually available online for many complex disease and traits. Second, our approach explores the functional association of chemicals and diseases from the genomic perspective, thus the results should be more robust to overcome the shortcomings of traditional methods, such as it is difficult to accurately measure *in vivo* exposure.

The tool of CGSEA is mainly developed for scanning candidate chemicals associated with human complex diseases or traits. Due to the following two reasons, we only used the chemical related gene sets collecting from human and rice. First, many of the organisms included in CTD are invertebrates and non-mammal, such as cnidarians and ctenophores. Due to different genetic background, it is difficult to generalize the results to other organisms. Second, the mouse has a long and rich history in biological research, and many consider it a model organism for the study of human development and complex disease (Pennisi 2002, Bogue 2003). Therefore, we used the chemical related gene sets collecting from human and mice in this study.

However, two limitations of this approach should be noted. First, the performance of CGSEA may be affected by the accuracy of TWAS results and chemical related gene sets. Second, all subjects in this study are from European ancestry. Due to different genetic background, our study results should be interpreted with caution when applied to other populations.

ACKNOWLEDGMENTS

Please provide Acknowledgments

LITERATURE CITED

Bogue, C. W., 2003 Invited review: Functional genomics in the mouse: Powerful techniques for unraveling the basis of human development and disease. J. Appl. Physiol. 94: 2502–2509. https://doi.org/10.1152/japplphysiol.00209.2003

Chen, J., I. Alberts, and X. Li, 2014 Dysregulation of the IGF-I/PI3K/AKT/mTOR signaling pathway in autism spectrum disorders. Int. J. Dev. Neurosci. 35: 35–41. https://doi.org/10.1016/j.ijdevneu.2014.03.006

Choi, J. K., and S. C. Kim, 2007 Environmental Effects on Gene Expression Phenotype Have Regional Biases in the Human Genome. Genetics 175: 1607–1613. https://doi.org/10.1534/genetics.106.069047

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel *et al.*, 2009 Common regulatory variation impacts gene expression in a cell type dependent manner. Science 325: 1246–1250. https://doi.org/10.1126/science.1174148

Goto, Y., and A. A. Grace, 2006 Alterations in Medial Prefrontal Cortical Activity and Plasticity in Rats with Disruption of Cortical Development. Biol. Psychiatry 60: 1259–1267. https://doi.org/10.1016/j.biopsych.2006.05.046

Gusev, A., A. Ko, H. Shi, G. Bhatia, W. Chung *et al.*, 2016 Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. 48: 245–252. https://doi.org/10.1038/ng.3506

Kim, J., M. K. Kim, J. K. Lee, J. H. Kim, S. K. Son *et al.*, 2010 Intakes of Vitamin A, C, and E, and beta-Carotene Are Associated With Risk of Cervical Cancer: A Case-Control Study in Korea. Nutr. Cancer 62: 181–189. https://doi.org/10.1080/01635580903305326

Mattingly, C. J., G. T. Colby, M. C. Rosenstein, J. N. Forrest, and J. L. Boyer, 2004 Promoting comparative molecular studies in environmental health research: an overview of the comparative toxicogenomics database (CTD). Pharmacogenomics J. 4: 5–8. https://doi.org/10.1038/sj.tpj.6500225

Mattingly, C. J., M. C. Rosenstein, A. P. Davis, G. T. Colby, J. N. Forrest, Jr. *et al.*, 2006 The Comparative Toxicogenomics Database: A Cross-Species Resource for Building Chemical-Gene Interaction Networks. Toxicol. Sci. 92: 587–595. https://doi.org/10.1093/toxsci/kfl008

Mccarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.*, 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9: 356–369. https://doi.org/10.1038/nrg2344

McHale, C. M., L. Zhang, Q. Lan, G. Li, A. Hubbard *et al.*, 2009 Changes in the peripheral blood transcriptome associated with occupational benzene exposure identified by cross-comparison on two microarray platforms. Genomics 93: 343–349. https://doi.org/10.1016/j.ygeno.2008.12.006

Navasacien, A., E. K. Silbergeld, R. Streeter, J. M. Clark, T. A. Burke *et al.*, 2005 Arsenic Exposure and Type 2 Diabetes: A Systematic Review of the Experimental and Epidemiologic Evidence. Environ. Health Perspect. 5: 641–648.

Olden, K., and S. Wilson, 2000 Environmental health and genomics: visions and implications. Nat. Rev. Genet. 1: 149–153. https://doi.org/10.1038/35038586

Papathanasiou, K., C. Gianoulis, A. Tolikas, D. Dovas, J. Koutsos *et al.*, 2005 Effect of depleted uranium weapons used in the Balkan war on the incidence of cervical intraepithelial neoplasia (CIN) and invasive cancer of the cervix in Greece. Clin. Exp. Obstet. Gynecol. 1: 58–60.

Pennisi, E., 2002 Genomics. Sequence Tells Mouse, Human Genome Secrets. Science 298: 1863–1865. https://doi.org/10.1126/science.298.5600.1863

Smith, M. T., 2010 Advances in Understanding Benzene Health Effects and Susceptibility. Annu. Rev. Public Health 31: 133–148. https://doi.org/10.1146/annurev.publhealth.012809.103646

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.*, 2005 Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102: 15545–15550. https://doi.org/10.1073/pnas.0506580102

Wang, K., M. Li, and M. Bucan, 2007 Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am. J. Hum. Genet. 81: 1278–1283. https://doi.org/10.1086/522374

Wen, Y., W. Wang, X. Guo, and F. Zhang, 2016 PAPA: a flexible tool for identifying pleiotropic pathways using genome-wide association study summaries. Bioinformatics 32: 946–948. https://doi.org/10.1093/bioinformatics/btv668

Zhou, X., P. Carbonetto, and M. Stephens, 2013 Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet. 9: e1003264. https://doi.org/10.1371/journal.pgen.1003264

Communicating editor: T. Matise