ERGR: An ethanol-related gene resource

An-Yuan Guo1, Bradley T. Webb2, Michael F. Miles3,4, Mark P. Zimmerman4, Kenneth S. Kendler1,5 and Zhongming Zhao1,4,5,*

1Department of Psychiatry and Virginia Institute for Psychiatric and Behavioral Genetics, 2Department of Pharmacy, 3Department of Pharmacology/Toxicology, Virginia Commonwealth University, Richmond, VA 23298, 4Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284 and 5Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA 23298, USA

ABSTRACT
Over the last decade rapid progress has been made in the study of ethanol-related traits including alcohol abuse and dependence, and behavioral responses to ethanol in both humans and animal models. To collect, curate, integrate these results so as to make them easily accessible and interpretable for researchers, we developed ERGR, a comprehensive ethanol-related gene resource. We collected and curated more than 30 large-scale data sets including linkage, association and microarray gene expression from the literature and 21 mouse QTLs from public databases. At present, the ERGR deposits ethanol-related information of ~7000 genes from five organisms: human (3311), mouse (2129), rat (679), fly (614) and worm (228). ERGR provides gene annotations and orthologs, detailed gene study information (e.g. fold changes of gene expression, P-values), and both the text and BLAST searches. Moreover, ERGR has data integration tools such as for data union and intersection, and candidate gene selection based on evidence in multiple datasets or organisms. The ERGR database is evolving with new data releases. More functions will also be added. ERGR has a user-friendly web interface with browse and search functions at multiple levels. It is freely available at http://bioinfo.vipbg.vcu.edu/ERGR/.

INTRODUCTION
Alcohol dependence, ethanol response and ethanol-related traits have been extensively studied in both humans and animal models. It is now clear that there are correlations between acute behavioral responses to ethanol and ethanol consumption or incidence of alcoholism in both animals and humans (1). In humans, alcoholism (alcohol dependence) is a common, genetically influenced complex disorder across the world. Family, twin and adoption studies demonstrated that genetic factors play a strong role in the etiology of alcoholism, accounting for 50–60% of the population variance in both men and women (2,3). Although genetic factors are important, alcoholism is a complex disease with environmental influences. Further, the architecture likely involves many genes with small effects along with environmental influences, as well as potential interactions between them. Therefore, it is a challenge to explore the molecular mechanisms underlying the genetic propensity to excessive alcohol consumption and use these for the development of new treatments for alcoholism. Many experimental strategies [linkage scan, association study, quantitative trait loci (QTLs) and microarray gene expression] have been applied in the studies of alcoholism and ethanol response in order to identify genes or chromosomal regions in both humans and model organisms (4,5).

Rapid progress in genetic studies over the past decade has identified a relatively large number of chromosomal locations or candidate genes that are linked to alcoholism, alcohol-related phenotypes and behavioral responses to ethanol (6–13). Human genetic studies have generally focused on alcohol dependence using linkage studies and association studies. The recent advances of high-throughput molecular technologies such as large-scale genotyping and DNA microarrays have greatly accelerated the generation of data used in studies searching for specific variants contributing to the genetic risk for alcoholism and ethanol response behaviors (14,15). The increasing rate of production for ethanol-related data is expected to accelerate in the near future since the cost of conducting genome-wide association studies (GWAS) is decreasing rapidly. Thus, these data provide us an unprecedented opportunity for integrating and making the wealth of results easily accessible and interpretable. So far, a few databases and computational tools, such as

*To whom correspondence should be addressed. Tel: +1 804 828 8129; Fax: +1 804 828 1471; Email: zzhao@vcu.edu
© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Here, we present such a database of ethanol-related gene resource (ERGR, http://bioinfo.vipbg.vcu.edu/ERGR/). To the best of our knowledge, it is a unique public database for ethanol-related genes. Aiming to efficiently integrate and analyze all or most of the published ethanol-related gene studies, we collected and annotated the representative large-scale ethanol-related gene datasets. These data were generated by different approaches including linkage scan, genome-wide association study, microarray expression, QTLs, retrieved from other public databases, or collected by a systematic literature search. We obtained data from the five most-studied model organisms: human, mouse, rat, fly and worm. In addition to information such as dataset description, gene annotations and gene ortholog information, the ERGR also provides tools for data integration (e.g., data union and intersection) and candidate gene selection based on multiple datasets or organisms. ERGR seeks to be a useful resource for the ethanol research community and a model database of data collection and integration for other complex diseases such as schizophrenia and Alzheimer’s disease.

DATA SOURCE AND METHODS

Data collection and curation

Currently, ERGR contains ethanol-related gene data from five organisms (human, mouse, rat, fly and worm). These data were collected from different technology platforms that have been widely used in alcohol dependence or ethanol response studies: linkage scan, genome-wide association study, microarray gene expression and QTLs, or by literature search. We collected these data by the following three approaches. First, we searched publications of large-scale alcohol dependence or ethanol response studies in NCBI PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) and then extracted and checked the data from these publications. Using this approach, we collected data including alcohol-related microarray gene expression studies from human or other animals, human alcohol dependence linkage studies and genome-wide association study. Specifically for the linkage data, we selected linkage regions by LOD scores and obtained the physical locations of the corresponding markers from UCSC genome browser (http://genome.ucsc.edu/) (19). Then, we retrieved the genes in the linkage regions from the Ensembl database (http://www.ensembl.org/index.html). Besides the genes in alcoholism or ethanol response studies, we also included other related genes. For example, we collected an addiction candidate gene list in which 130 genes were selected for a haplotype-based analysis of addiction (20).

Second, we extracted related data from other public databases. We obtained the mouse alcohol behavior related QTLs from the PARC (Portland Alcohol Research Center, http://www.ohsu.edu/parc/by_phen.shtml). Only those QTLs that were marked significant and whose genomic locations could be identified were extracted. Then, we retrieved genes mapped in the QTL regions from the Mouse Genome Informatics (MGI, http://www.informatics.jax.org/) (21). Further, we searched alcohol-related genes in HuGE Navigator (http://www.hugenavigator.net/), a database of genetic associations and human genome epidemiology (22). We found nine alcohol-related phenotypes in HuGE Navigator and extracted all the genes associated with them.

Third, we performed a systematic literature search by searching the titles and abstracts of all the publications available in PubMed. We searched each protein coding gene symbol with one of the three keywords: alcohol, ethanol and alcoholism. To reduce false-positives, we manually checked those gene symbols having fewer than three letters/digits or having more than 100 hits of publications.

For each dataset, we compiled a summary description from the original data source. The summary includes the experimental method and treatment, platform, organism, tissue and phenotype, and the publication. For each gene in a dataset, more detailed information was extracted from the data source such as gene expression fold change, P-value, and tissue type.

Gene ID

We used NCBI Entrez Gene ID as the central ID for cross linking and annotation. However, specific studies used different IDs, such as gene symbol, mRNA accession number, EST accession number, clone ID, Affymetrix probe ID, Ensembl ID or UniGene ID. We applied the following three approaches to convert different IDs to gene IDs. (i) We downloaded gene2accession, gene2unigene, and gene_info files from NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/) and obtained the corresponding gene IDs for the accession numbers, UniGene IDs or gene symbols used in different studies. (ii) We used an online tool, IDConverter (http://idconverter.bioinfo.cnci.es/IDConverter.php), to convert the original IDs in the publications to gene IDs (23). (iii) When the IDs could not be converted by the above two approaches, we manually searched the NCBI databases for the gene IDs.

Gene annotation

We downloaded gene annotation files from the NCBI FTP site. Then, we extracted annotation information for the genes in our database from these files. We parsed the gene_info file to retrieve the gene information such as gene symbol, alias, full name, chromosome, genetic location and gene type. We obtained gene ontology (GO) annotations from the gene2go file downloaded from the GO website (http://www.geneontology.org/) (24). The accession numbers of the reference sequences and chromosomal location of each gene was parsed from the gene2refseq file.

Orthologs

A cross-species gene discovery and validation scheme can provide both powerful confirmation of candidate genes
and mechanistic information about gene-behavior relationships. Thus, we searched orthologs of the ethanol-related genes in this database. For the nonhuman genes in our datasets, their human ortholog information was obtained and curated. We obtained human/mouse and human/rat ortholog information from MGI (ftp://ftp.informatics.jax.org/pub/reports/index.html), and human/fly and human/worm from the Inparanoid database (version 6.1, http://inparanoid.sbc.su.se/download/6.1/) (25).

DATABASE CONTENT AND ORGANIZATION

Data overview
As summarized in Table 1, we have collected and curated more than 30 genome-wide or large-scale datasets including microarray gene expression, linkage and genome-wide association studies, results of literature search and 21 mouse QTLs. At present, ERGR includes ∼7000 genes in five organisms: human (3311), mouse (2129), rat (679), fly (614) and worm (228). For rat, fly and worm, ERGR only has microarray expression data. Because of the major research interest in humans, the human data in ERGR is the most comprehensive, which includes microarray gene expression, linkage studies, genome-wide association study and literature search results.

Database organization
We used MySQL, a SQL client/server relational open source database management system that has been commonly used in the development of biomedical databases, to store and manage the data. One table was specifically designed to store the summary descriptions of all datasets. Because the formats of the datasets generated by the different methods often varied, we managed datasets of gene expression, linkage, association or QTLs by separate tables. Annotations of gene information, gene ontology, reference sequences and ortholog information were also stored in individual tables. Dataset name, PubMed ID and Entrez gene ID are the keys to link between tables.

Table 1. Summary of the data in ERGR

| Species          | Method                  | No. of datasets | No. of genes | Source                          |
|------------------|-------------------------|-----------------|--------------|---------------------------------|
| Human (Total: 3311 genes) | Microarray expression 7  | 831             | Literature   |
|                  | Genome-wide association 1 | 71              | Literature   |
|                  | Linkage 3               | 918             | Literature   |
|                  | HuGE Navigator 9 diseases | 203            | HuGE Navigator (http://www.hugenavigator.net/) |
|                  | Addiction array list 1  | 130             | Literature   |
|                  | Literature search 1     | 1726            | Literature   |
| Mouse (Total: 2129 genes)  | Microarray expression 11 | 682             | Literature   |
|                  | QTL 21 QTLs            | 1568            | PARC Alcohol QTLs (http://www.ohsu.edu/parc/by_phen.shtml) |
| Rat              | Microarray expression 6 | 679             | Literature   |
| Fly              | Microarray expression 2 | 614             | Literature   |
| Worm             | Microarray expression 1 | 228             | Literature   |

WEB INTERFACE

A user-friendly web interface was designed and implemented for ERGR. It is freely available at http://bioinfo.vipbg.vcu.edu/ERGR/. The user can browse and search all the data at different levels or combine the data by the integration functions.

Data browse
To help the user to browse the data easily, ERGR provides four different browsing methods: (i) by species; (ii) by method or platform such as microarray expression, linkage or association; (iii) by chromosome of a species; (iv) or through a summary page which lists all the datasets available in ERGR. A cascading style is applied for dataset browse, i.e., from dataset list to gene list, and then to gene information. By clicking the dataset name on a dataset list page it will show the dataset description and the corresponding list of genes. Selecting the gene ID will link to the gene information page, which includes gene ID, symbol and name, GO annotation, RefSeq, chromosome location, database cross links and the detailed study information (e.g. fold change of gene expression and $P$-value) extracted from the original ethanol studies. For example, based on the microarray dataset ‘15816859’, the fold change and $P$-value of ADH6 gene expression were 0.69 and $7.00 \times 10^{-3}$, respectively (http://bioinfo.vipbg.vcu.edu/ERGR/geneinfo.php?id=130). Moreover, the user may find the detailed information of the studied single nucleotide polymorphisms (SNPs) via dynamic links to the NCBI dbSNP database.

Data search
ERGR provides three approaches for searching the data including text search and sequence search. First, the user may find a quick search box in the top right of the web page for searching gene ID and symbol. It supports wildcard searches such as using partial gene symbol (e.g. ADH) or using an asterisk (e.g. ADH*). Second, ERGR provides an advanced search page, on which users may combine different search terms (e.g. ID, symbol, phenotype, physical location and GO term) for a user-defined search.
Third, BLAST search against the nucleotide or protein sequences of the ethanol-related genes in each organism or all the five organisms is available in ERGR.

DATA INTEGRATION

One current opportunity and challenge is the increasing large amount of public data that can be applied to the study of alcohol-related traits. Given the growth and scale of these data, efficient integration is necessary in studying a complex disorder such as alcohol dependence. Currently, ERGR provides the users data union and intersection functions for data integration; however, more functions are being developed in this ongoing project. Moreover, ERGR provides candidate gene selection results based on the evidence in multiple datasets and multiple organisms.
Data union and intersection

ERGR supports the union operation of any two datasets from the same organism and the intersection operation of any two datasets. There are three rules for the dataset intersection operation. First, if the two datasets to be compared are from the same organism, ERGR performs the intersection operation and outputs the results based on gene ID. Second, if the two datasets are from different organisms and one of them is from the human, ERGR uses human genes as the reference, transforms non-human genes to the human orthologs, and then performs the intersection operation based on human gene IDs. Third, if the two datasets are from two different non-human organisms, ERGR transforms genes in the both datasets to human orthologs and then performs the intersection operation based on human IDs. An example of data integration is shown in Figure 1.

Candidate gene selection

The candidate gene selection and prioritization function can be used to select candidate genes for follow up experimental replication or bioinformatics analysis. To make the ERGR data more useful and serve the community more effectively, ERGR provides some candidate gene selection results based on the evidence in multiple datasets either in one organism or multiple organisms. At present, ERGR contains four such candidate gene lists for ethanol response or alcoholism-related traits. The first one is a candidate gene list generated from the datasets of all organisms using human genes as reference. The non-human genes were mapped to human orthologous genes so that they could be compared systematically. There were 42 human genes or their orthologs that had evidence in more than four datasets. The other three candidate gene lists were selected from all the available datasets in one organism (human, mouse or rat) only. These candidate genes include those that have been well studied for alcohol dependence such as ADH, ALDH, GABA receptors, and NPY (4). However, the candidate gene lists also include genes with multiple lines of evidence but not so well studied such as CPE, GFAP, CRYAB, GAD1 and NTRK2, among which two [GAD1 (26) and NTRK2 (27)] had association studies reported. We found three genes (KCNJ9, GNB1 and ATP1A2) that had evidence in at least five datasets from the mouse, rat or fly but no evidence yet in any human datasets.

FUTURE PERSPECTIVES

ERGR is a unique database for ethanol-related genes and their annotations. It is freely available to public and also serves as a core data management system for the local VCU alcohol research community (VCU ARC). We will continue to collect and curate ethanol-related data, especially the genome-wide association studies. We will develop more tools that allow the users to customize their gene ranking, track their own data, and present the genes or integration results graphically.
frontal and motor cortices of human alcoholics. J. Neurochem., 81, 802–813.
14. Hoffman, P. and Tabakoff, B. (2005) Gene expression in animals with different acute responses to ethanol. Addict. Biol., 10, 63–69.
15. Johnson, C., Drgon, T., Liu, Q.R., Walther, D., Edenberg, H., Rice, J., Foroud, T. and Uhl, G.R. (2006) Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. Am. J. Med. Genet. B Neuropsychiatr. Genet., 141B, 844–853.
16. Wang, J., Williams, R.W. and Manly, K.F. (2003) WebQTTL: web-based complex trait analysis. Neuroinformatics, 1, 299–308.
17. Bhave, S.V., Hornbaker, C., Phang, T.L., Saba, L., Lapadat, R., Kechris, K., Gaydos, J., McGoldrick, D., Dolbey, A., Leach, S. et al. (2007) The PhenoGen informatics website: tools for analyses of complex traits. BMC Genet., 8, 59.
18. Zhang, B., Kirov, S. and Snoddy, J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res., 33, W741–W748.
19. Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F. et al. (2008) The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res., 36, D773–D779.
20. Hodgkinson, C.A., Yuan, Q., Xu, K., Shen, P.H., Heinz, E., Lobos, E.A., Binder, E.B., Cubells, J., Ehlers, C.L., Gelernter, J. et al. (2008) Addictions Biology: Haplotype-Based Analysis for 130 Candidate Genes on a Single Array. Alcohol Alcohol., 43, 505–515.
21. Bult, C.J., Eppig, J.T., Kadon, J.A., Richardson, J.E. and Blake, J.A. (2006) The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res., 36, D724–D728.
22. Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. and Khoury, M.J. (2008) A navigator for human genome epidemiology. Nat. Genet., 40, 124–125.
23. Alibes, A., Yankilevich, P., Canada, A. and Diaz-Uriarte, R. (2007) DIconverter and IDClight: conversion and annotation of gene and protein IDs. BMC Bioinformatics, 8, 9.
24. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25, 25–29.
25. Berglund, A.C., Sjolund, E., Ostlund, G. and Sonnhammer, E.L. (2008) InParanoid 6: eukaryotic ortholog clusters with paralogs. Nucleic Acids Res., 36, D263–D266.
26. Loh, W., Lane, H.Y., Chen, C.H., Chang, P.S., Ku, L.W., Wang, K.H. and Cheng, A.T. (2006) Glutamate decarboxylase genes and alcoholism in Han Taiwanese men. Alcohol. Clin. Exp. Res., 30, 1817–1823.
27. Xu, K., Anderson, T.R., Neyer, K.M., Lamparella, N., Jenkins, G., Zhou, Z., Yuan, Q., Virkkunen, M. and Lipsky, R.H. (2007) Nucleotide sequence variation within the human tyrosine kinase B neurotrophin receptor gene: association with antisocial alcohol dependence. Pharmacogenomics J., 7, 368–379.