AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology® Resource Announcements

COnTORT: COmprehensive Transcriptomic ORganizational Tool for Simultaneously Retrieving and Organizing Numerous Gene Expression Data Sets from the NCBI Gene Expression Omnibus Database

Kevin S. Myers,[a] Michael Place,[a] Daniel R. Noguera,[a,b] Timothy J. Donohue[a,c]

[a]Wisconsin Energy Institute and Great Lakes Bioenergy Research Center, University of Wisconsin—Madison, Madison, Wisconsin, USA
[b]Department of Civil & Environmental Engineering, University of Wisconsin—Madison, Madison, Wisconsin, USA
[c]Department of Bacteriology, University of Wisconsin—Madison, Madison, Wisconsin, USA

ABSTRACT We introduce COnTORT (**CO**mprehensive **T**ranscriptomic **OR**ganizational **T**ool), a publicly available program that retrieves all available gene expression data and associated metadata for an organism from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. The data are compiled into text files that can be used for downstream bioinformatic applications.

The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database is a centralized repository for millions of functional genomic data sets, including gene expression data from both microarrays and high-throughput sequencing (1–3). While there are tools from both the NCBI and others (e.g., GEOquery [4] and GEOparse [https://geoparse.readthedocs.io]) to allow for access to individual data sets, there is no publicly available tool that will allow users to bulk-retrieve GEO data and organize all the available gene expression data for a given organism. Overcoming this limitation will increase the use of NCBI GEO data for large-scale analyses using the available gene expression data. Here, we introduce COnTORT (**CO**mprehensive **T**ranscriptomic **OR**ganizational **T**ool), a publicly available Python3 program that facilitates the retrieval of all publicly available gene expression data derived from a search in the NCBI GEO. COnTORT requires only a GenBank gene annotation file for the organism of choice (5) and will perform a search of the NCBI GEO entered by the user upon running the program (Fig. 1). This query can be as specific or general as desired, meaning the results can be narrow or broad to answer a particular question.

COnTORT downloads all NCBI GEO data associated with the search term provided and uses these data to produce tab-delimited text files containing the organized results of the gene expression data. The first five columns contain the annotation information from the GenBank file (locus tag, old locus tag, gene name, gene synonyms, and product), and the remaining columns contain the gene expression data for each experimental sample. One output file will mean-center the data based on experiments to control for differences between experiments, while the second output file will not mean-center the data. These output files are in a format that can be used for virtually any downstream bioinformatic application.

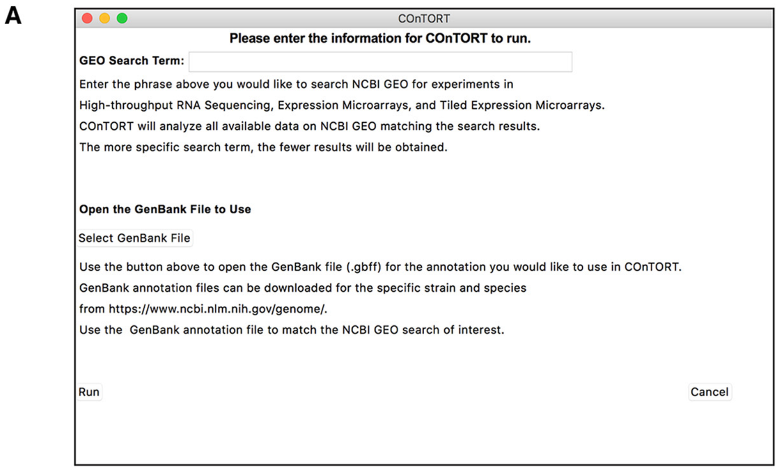
Acquiring numerous bacterial samples takes about 20 min, while retrieving data from yeast or humans, with much larger genomes, takes around an hour to complete on a standard desktop computer. COnTORT thus provides a significant savings in time and effort for researchers in gaining access to large numbers of gene expression data sets from a single organism. We have used COnTORT files to correlate expression of genes across a large number of diverse data sets to identify new targets of transcription

Citation Myers KS, Place M, Noguera DR, Donohue TJ. 2020. COnTORT: **CO**mprehensive **T**ranscriptomic **OR**ganizational **T**ool for simultaneously retrieving and organizing numerous gene expression data sets from the NCBI Gene Expression Omnibus database. Microbiol Resour Announc 9:e00587-20. https://doi.org/10.1128/MRA.00587-20.

Editor Catherine Putonti, Loyola University Chicago

Address correspondence to Timothy J. Donohue, tdonohue@bact.wisc.edu.

Received 20 May 2020
Accepted 24 May 2020
Published 18 June 2020



Please enter the information for COnTORT to run.

GEO Search Term:

Enter the phrase above you would like to search NCBI GEO for experiments in
High-throughput RNA Sequencing, Expression Microarrays, and Tiled Expression Microarrays.
COnTORT will analyze all available data on NCBI GEO matching the search results.
The more specific search term, the fewer results will be obtained.

Open the GenBank File to Use

Use the button above to open the GenBank file (.gbff) for the annotation you would like to use in COnTORT.
GenBank annotation files can be downloaded for the specific strain and species
from https://www.ncbi.nlm.nih.gov/genome/.
Use the GenBank annotation file to match the NCBI GEO search of interest.

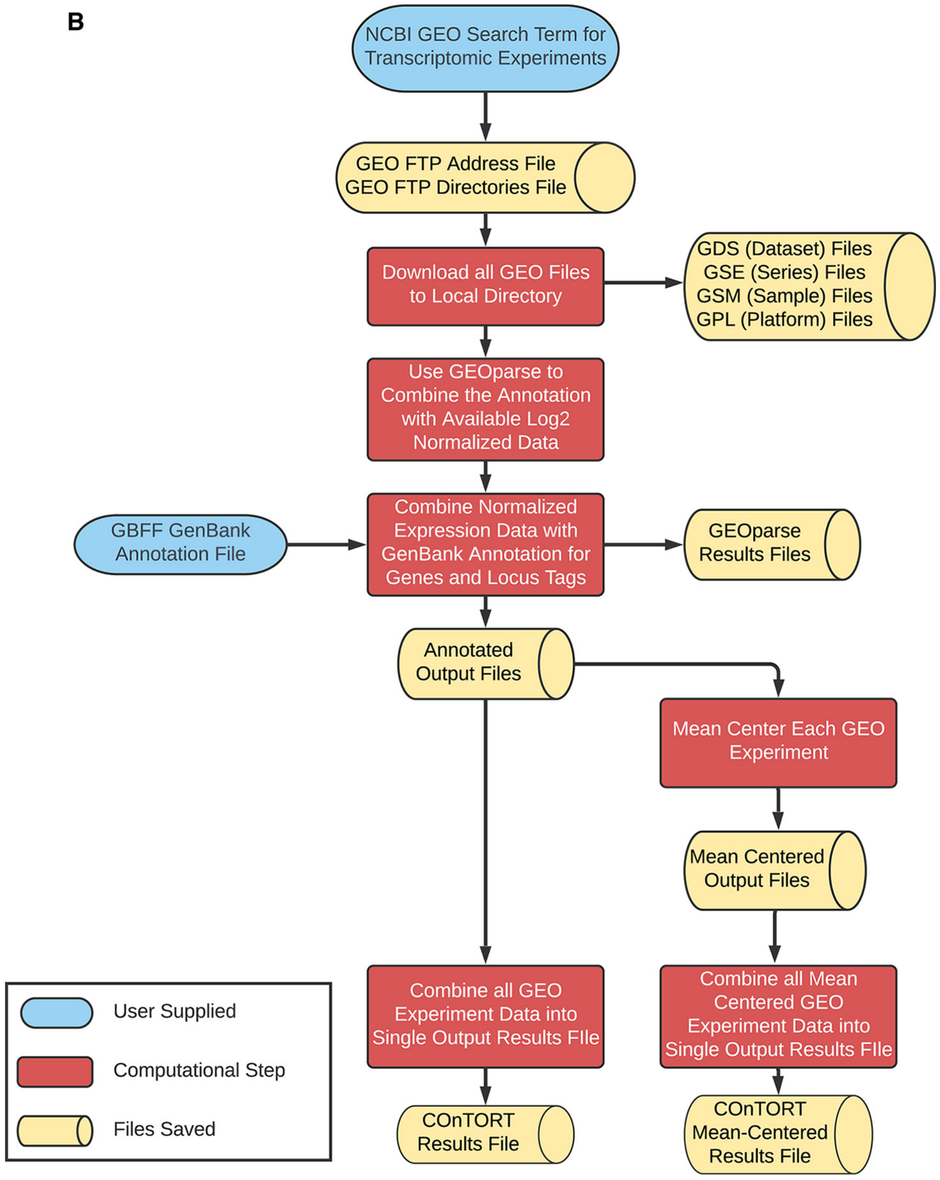

B
NCBI GEO Search Term for Transcriptomic Experiments
GEO FTP Address File
GEO FTP Directories File
Download all GEO Files to Local Directory
GDS (Dataset) Files
GSE (Series) Files
GSM (Sample) Files
GPL (Platform) Files
Use GEOparse to Combine the Annotation with Available Log2 Normalized Data
GBFF GenBank Annotation File
Combine Normalized Expression Data with GenBank Annotation for Genes and Locus Tags
GEOparse Results Files
Annotated Output Files
Mean Center Each GEO Experiment
Mean Centered Output Files
Combine all GEO Experiment Data into Single Output Results FIle
Combine all Mean Centered GEO Experiment Data into Single Output Results FIle
COnTORT Results File
COnTORT Mean-Centered Results File
User Supplied
Computational Step
Files Saved

FIG 1 Overview of the COnTORT process. (A) COnTORT interface where the user enters the query to search NCBI GEO and selects the GenBank file to be used. (B) Flowchart of the COnTORT program. NCBI GEO search

(Continued on next page)

FIG 1 Legend (Continued)
results are parsed, and all GEO data are downloaded. Normalized gene expression data from the NCBI GEO are retrieved. The normalized files are organized with the annotation from the GenBank file, mean-centered, and combined into COnTORT result files.

factors in bacteria. While COnTORT builds upon critically important tools such as GEOparse and GEOquery, and indeed uses GEOparse, it allows researchers to harness the power of the millions of gene expression data sets in the NCBI GEO database.

COnTORT is written for the analysis of gene expression data. However, because the scripts are publicly available, changes can be made to allow COnTORT to download and combine all the available data from chromatin immunoprecipitation (ChIP)-chip or ChIP-seq experiments. Further, COnTORT can be modified to acquire data in any online database, such as those housed by the European Bioinformatics Institute (6). While COnTORT is a valuable tool for the analysis of gene expression data, it also provides a platform on which additional tools and analyses can be built.

Data availability. COnTORT is available at https://github.com/GLBRC/contort and https://pypi.org/project/contort. A detailed description and tutorial are available at https://github.com/GLBRC/contort/blob/master/Additional_Information.md.

ACKNOWLEDGMENTS

We thank the people who helped by commenting on the manuscript and testing the COnTORT software.

This material is based upon work supported by the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under award number DE-SC0018409.

REFERENCES

1. Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210. https://doi.org/10.1093/nar/30.1.207.
2. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. 2013. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995. https://doi.org/10.1093/nar/gks1193.
3. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. 2007. NCBI GEO: mining tens of millions of expression profiles—database and tools update. Nucleic Acids Res 35:D760–D765. https://doi.org/10.1093/nar/gkl887.
4. Davis S, Meltzer PS. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23:1846–1847. https://doi.org/10.1093/bioinformatics/btm254.
5. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019. GenBank. Nucleic Acids Res 47:D94–D99. https://doi.org/10.1093/nar/gky989.
6. Brooksbank C, Cameron G, Thornton J. 2010. The European Bioinformatics Institute's data resources. Nucleic Acids Res 38:D17–D25. https://doi.org/10.1093/nar/gkp986.