NetGO: A rapid visualization and reference tool for gene expression data

Jonathan Gerhart,[1,2,3] Bridget Neary,[1] Will Hutwagner,[1] Mani Jain,[1] Joey Gonzales-Donez[1]

1: Georgia Institute of Technology 2: Centers for Disease Control and Prevention, Division of Parasitic Disease and Malaria 3: Author for correspondence. Contact via nwx7@cdc.gov

In collaboration with VectorBase

Motivation

- Advances in Next Generation Sequencing (NGS) has made it possible for research laboratories to rapidly and accurately measure the transcription of DNA in an organism on demand through RNA-seq.
- While NGS technology has seen widespread adoption in academic, corporate and public health sectors, a dedicated staff is often required to collect, analyze, interpret and translate the information into data that can be used by a sponsor to make decisions of medical or financial impact.
- While many programs exist to provide in-depth, accurate and exhaustive RNA-seq analysis, surveys of many public health scientists report that these tools are unintuitive, opaque or overly technical
- To address, NetGO can allow for a quick-and-dirty instant analysis of trends in NGS data, empowering users to more efficiently direct efforts in research or validation.

Data Sources and Collaboration

For our initial trials, we partnered with VectorBase, a comprehensive resource for bioinformatics information on vectors of human diseases.
While designing NetGO, we contributed expertise to the Vectorbase Project, in the form of a supplementary project consisting of improvements to database integration, and population visualization tools to better understand how to implement NetGO to fill needs in the bioinformatic community.

Figure 1: Improvements to VectorBase Population Browser

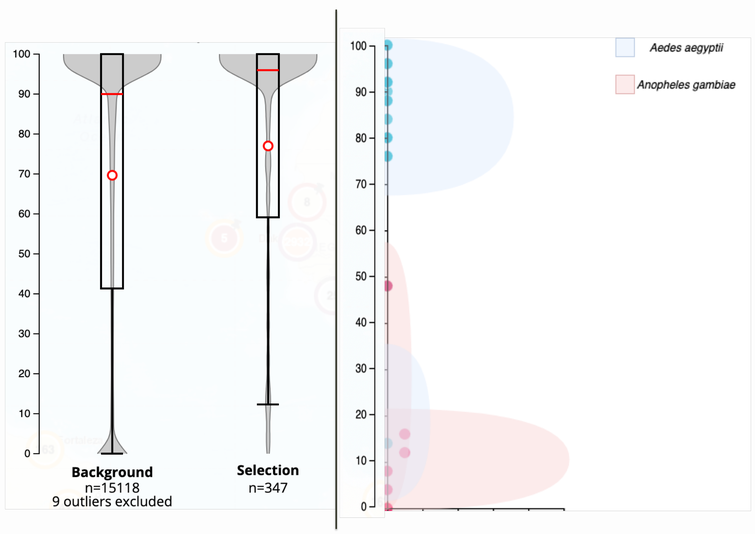
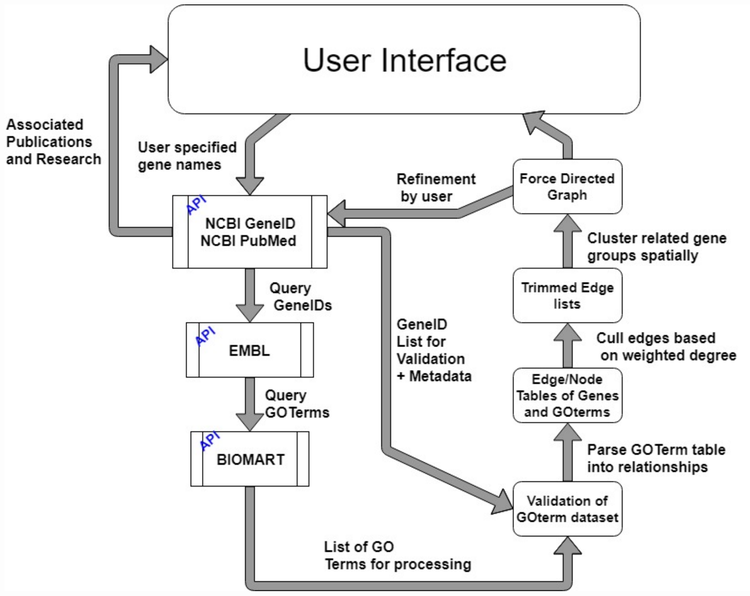


Figure 1: Vectorbase utilizes an in-house modification of Esri Leaflet to visualize population level data (left). As part of our initial research into effective communication of bioinformatic data, we refined the population browser to better support simultaneous visualization of groups of distinct phenotypes and metrics by converting the standard single violin plot to a more intuitive set of overlaid plots and a simple mouse-over filtering system (right)

Figure 2: Structure and execution of NetGO



NetGO uses a series of API calls to reference databases to collect and collate data on gene sets, before parsing that information into networks of related genes and presenting them to the user for review and refinement.

Approach

Design summary:

- Collect over-expressed gene lists from the user and cross reference against entries in standard formats against a wide range of internationally respected databases.
- Collate and parse the gathered data into an easy to digest report using a simple and lightweight text comparison and sanity checking algorithm.
- Parse the functional Gene ontology (GO) terms collected from these databases into a network.
- Utilize edge regression to optimize paths between genes via gene ontology network, rapidly building networks of related genes based on official annotations.
- Present the user with an interactive force directed graph that intuitively shows the relationship between different clusters of genes of interest.

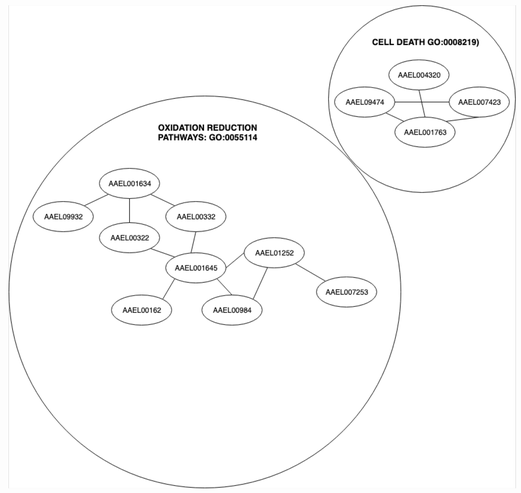
Justification:

Most of the method described mirrors a traditional RNASeq annotation and description pipeline. To promote accessibility, a simple interface and succinct presentation through the force directed graph was selected.

Novelty:

While many elements described here have been replicated in other popular tools, our approach and focus is novel. This is a strength of NetGO, as our focus on interpretability and accessibility offers advantages and occupies a unique niche on top a foundation of accepted methods and databases.

Figure 3: Current UI and Implementation



NetGO's UI output shown clustering a selection of genes in a simulated trial. The assigned GO Term corresponds to the lowest common linking term shared between every element of the network.

Experiments

We experimented with several visualization and gene organization strategies in development of NetGO.

We conducted user surveys of different organization schemes for gene expression data, including mimicking biochemical assays in a simulated microarray, visualizing the Gene Ontology terms in a collapsible tree, and the inclusion of other reference data directly on the network.

We presented all of these methods to trained biologists with minimal bioinformatics experience, and reviewed qualitative feedback for accuracy, information recapitulation and user reported comfort.

Our clustering algorithm was scored in a similar method. Clusters from different weights, forcing parameters and metrics where reviewed by biologists for their ability to cluster known gene networks together.

Improvements to VectorBase Databases

During initial development of our backend for NetGO, we developed a reference system to improve VectorBase's community annotation service. Prior, VectorBase had been providing information parsed from a manual download of the National Center for Biotechnology Information (NCBI) database of genes in 2017. We tested our database cross-referencing methods by building a small script to automatically curate the data hosted on vecotrbase servers, using a simple web crawler and text parser, paired with NCBI's api to identify depreciated links and out of date descriptions.

ACKNOWLEDGEMENTS

We thank Bob McCallum, Samuel Rund, Mary Ann McDowell and Daniel Larson from the vectorbase team for their collaboration and interest. Thanks also to Audrey Lenhart, Seth Irish, Nsa Dada and Lucy Impoinvil from CDC Entomology Branch for their input and patience. We also thank Isabela Arenta for her participation in the initial design stages.