METHOD ARTICLE
Introducing high school students to the Gene Ontology classification system [version 1; referees: awaiting peer review]

Mehek Dedhia¹, Kenneth Kohetuk ², Wim E. Crusio ³,⁴, Anna Delprato ¹

¹BioScience Project, Wakefield, MA, 01880, USA
²Saint Dominic Savio Catholic High School, Austin, TX, 78717, USA
³Institut de Neurosciences Cognitives et Intégratives d’Aquitaine (UMR 5287), Pessac, France
⁴University of Bordeaux (UMR 5287), Pessac, France

Abstract
We present an activity that introduces high school students to the Gene Ontology classification system which is widely used in genomics and systems biology studies to characterize large sets of genes based on functional and structural information. This is a valuable and standardized method used to identify genes that act in similar processes and pathways and also to gain insight into the overall architecture and distribution of genes and gene families associated with a particular tissue or disease. Through this exercise, students will learn how the classification system works by analyzing a list of genes using DAVID the Database for Annotation, Visualization and Integrated Discovery that incorporates the Gene Ontology system into its suite of analysis tools. This method of profiling genes is used by our high school student interns to categorize gene expression data related to behavioral neuroscience. Students will get a feel for working with genes and gene sets, gain vocabulary, obtain an understanding of how a database is structured and gain an awareness of the vast amount of information that is known about genes as well as the online analysis tools that are available.

Keywords
gene ontology, high school students, genomics

This article is included in the Teaching and communicating science in a digital age collection.
Introduction
Genomics is the branch of biology concerned with the study of genes and their functions (see the National Institutes of Health Frequently Asked Questions about Genetic and Genomic Science). Genomics arose from the acceleration of genetic research which was fueled by the development of rapid and affordable DNA sequencing technologies (Shendure et al., 2017). This opened the door to the sequating of entire genomes. Presently, the DNA codes for thousands of genomes from diverse species have been sequenced and studied (see the National Center for Biotechnology Information Genome database).

The goals in genomics research are to address all genes and their inter-relationships in order to understand the combined influence on the function of an organism. With this new found knowledge of the staggering number of genes that make up an organism, the Gene Ontology (GO) classification system was created to organize genes by their similarities and differences (see Gene Ontology Consortium ‘About’ page). "Ontology" is not a commonly encountered term and there are several definitions that are related to philosophical concepts.

In the context of information science, as described here, “ontology” is concerned with the representation, formal naming and classification system with the purpose of describing the relationship categories and properties of the data.

This classification system provides the scientific community with a structured vocabulary for defining genes (Ashburner et al., 2000; du Plessis et al., 2011; Hastings, 2017; Thomas, 2017). GO terms are commonly used in most, if not all, databases and analysis tools relevant to bioinformatics, systems biology (Wanjek, 2011), and genomics studies (du Plessis et al., 2011). GO terms are species specific and are continuously revised and expanded as biological knowledge is obtained (Gaudet et al., 2017).

The importance of the GO term system becomes apparent when analyzing the organization of genomes and coding regions, the distribution of genes involved in specific processes and the conservation of genes across species (Gaudet et al., 2017). This classification system is also quite powerful when analyzing data from large scale gene expression studies (du Plessis et al., 2011) that consider co-expression data from specific tissues obtained under defined circumstances such as treatment with pharmaceutical agents, or with neurodevelopmental disorders, cancer, or diabetes as examples. GO terms are instrumental for understanding the functions of these genes.

Introducing GO terms and the gene classification system to high school students will bring them up to speed on a commonly used research tool in current genomics methods and expose them to the vast amounts of data that have been derived from genomics and systems biology studies.

In the subsequent sections we show an example of how to extract information about a gene from its associated GO terms and then provide instruction for a practical exercise which will enable students to profile a list of genes using GO terms in the bioinformatics resource DAVID, The Database for Annotation, Visualization and Integrated Discovery. This is a protocol that we teach to our high school student interns when they are evaluating gene expression data for their summer projects (Crusio et al., 2017, see BioScience Project student posters).

Procedure
Making sense of a gene
The overall structure of GO is hierarchical and is based on parent-child terms where the parent term is broader and child term is more specialized.

GO terms group genes according to 3 categories, each of which are considered a distinct ontology: Molecular Function (MF, molecular-level activities performed by gene products), Biological Process (BP, the larger processes, or biological programs accomplished by multiple molecular activities), and Cellular Component (CC, the locations relative to cellular structures in which a gene product performs a function).

As an example, consider the GO term classification for the RAB5A gene (Figure 1). RAB5A belongs to a family of genes called Rab GTPases that are key regulators of intracellular membrane trafficking. Rabbs are involved in the formation of transport vesicles and their fusion with membranes. They are enzymes and mediate their function by cycling between a GDP bound inactive and a GTP bound active state. Because of their fundamental and ubiquitous role, this family of genes are associated with many biological processes and diseases.

The GO term classification for the RAB5A gene gives:

**GOTERM_BP**: endocytosis, phagocytosis, small GTPase mediated signal transduction, blood coagulation, protein transport, regulation of endocytosis, synaptic vesicle recycling, viral RNA genome replication, early endosome to late endosome transport, positive regulation of exocytosis, regulation of endosome size, regulation of filopodium assembly, receptor internalization involved in canonical Wnt signaling pathway, regulation of synaptic vesicle exocytosis, regulation of autopagosome assembly

**GOTERM_CC**: ruffle, intracellular, cytoplasm, endosome, early endosome, cytosol, plasma membrane, synaptic vesicle, endosome membrane, actin cytoskeleton, endocytic vesicle, axon, dendrite, phagocytic vesicle membrane, somatodendritic compartment, melanosome, neuronal cell body, terminal bouton, axon terminus, membrane raft, phagocytic vesicle, extracellular exosome, cytoplasmic side of early endosome membrane.

**GOTERM_MF**: GTPase activity, protein binding, GTP binding, GDP binding

From the RAB5A related GO terms, we get the overall impression that this gene encodes an enzyme, that is involved in signaling, transport and vesicle dynamics and is associated with cell membranes. How do we arrive at this description?
In this example, the information obtained from the MF category is that the protein product of the RAB5A gene binds to guanine nucleotides: GTP and GDP (Guanosine tri and di phosphate, respectively) and that it is an enzyme. This is evident by the “GTPase activity” term. Whenever the suffix “ase” is used in the context of a gene or protein, it refers to an enzyme, something that catalyzes a chemical reaction. For the BP category, there are several terms associated with intracellular transport, signaling, and endocytosis. Finally, the terms associated with CC include endosome and endosome-like organelles (melanosomes, synaptic vesicles, phagocytic vesicles), as well as membrane structures (ruffles, rafts).

**Gene Profiling in DAVID**
DAVID is primarily a clustering program that groups genes based on different criteria related to GO terms. DAVID links to other databases that contain complementary information like The Gene Ontology. In this exercise, students will use the sample gene lists (DEMOLIST1 or DEMOLIST2) that are accessible from the DAVID database to see how the Gene Ontology classification partitions a set of genes based on GO Terms. Screenshots and videos are provided for step by step instruction. We also provide a video to instruct students on profiling a gene list in DAVID obtained from a random gene list generator.

### Protocol

**Screenshot 1 (Figure 2).** DAVID landing page. The start analysis link is accessed here and is circled in red in this image. (Video 1 (Delprato et al., 2019a))

**Screenshot 2 (Figure 3).** Submitting a gene list. Select either DEMOLIST 1 or DEMOLIST 2 (left panel). The identifier will come up automatically because this is a demonstration list. If you are submitting your own gene list then, the identifier will have to be specified from the dropdown menu (Video 2 (Delprato et al., 2019b)). Typically the identifier is the “Official Gene Symbol”.

Click “Gene List”, then “Submit List” (Video 1 (Delprato et al., 2019a)).

**Screenshot 3 (Figure 4).** Species selection. You will see a notice: “Multiple Species, have been Detected”. Highlight “Homo Sapiens” in the window, Select “Homo Sapiens” below the window (Example - DEMOLIST 1: 149 genes, highlighted in grey, left panel). Next, you will see the message “Submission Successful” (Video 1 (Delprato et al., 2019a)).

**Screenshot 4 (Figure 5).** Obtaining the results. Select “Functional Annotation Tool”, beneath the blue arrow. Next, select “Functional Annotation Table”, Bottom of the page (Video 1 (Delprato et al., 2019a)).

**Screenshot 5 (Figure 6).** Reading the output. The gene ID and the full gene name are shown in the blue bars above each entry. The GO Term BP (Biological Process), GO Term CC (Cellular Component), and GO Term MF (Molecular Function), terms are clickable descriptors and link to the Gene Ontology website. See above for a complete description of the GO categories (Video 1 (Delprato et al., 2019a)).

**Screenshot 6 (Figure 7).** Keyword search. When selecting terms for a keyword search, a more complete outcome is achieved if just a few letters are specified. For example, “neur” will capture terms both starting with neuro and neural (Video 1 (Delprato et al., 2019a)). DAVID output can be searched for genes related to other process and diseases as well. Have students evaluate the gene list based on their interest. They can identify genes related to a particular process. Students may work individually or in groups.

### Optional exercise

Students may wish to try this with their own gene lists. This online gene list generator will enable students to generate a...
Figure 2. David landing page. Screenshot of the DAVID landing page containing a brief description of the site and links for available tools. The “start analysis” button is circled in red and is located at the top left side of the page. This is the first step in submitting a geneset for profiling with GO Terms in DAVID.

Figure 3. Submitting a geneset. Screenshot of the page where a geneset can be submitted “Step1: Enter Gene List”, there are options to copy and paste a geneset or upload a file. There is also an option to use either of two sample lists provided by the DAVID site. Under the pull down menu, “Step 2 Select Identifier”, there are many types of identification designations for the same gene. For the sample lists, the identifier will come up automatically. When submitting a gene list from the geneset generator, the identifier is “Official Gene Symbol” which is used in most cases. If the identifier that you choose does not match the identifiers in the submitted geneset, you will receive a message stating this and an option to convert to the correct identifier.
Figure 4. **Species selection.** Screenshot of a successfully submitted geneset. Here it is necessary to select the species. In this instance “Homo sapiens” is highlighted along with the number of genes that are recognized by the site and for which information is available. Once the species is highlighted, it is necessary to click the “Select Species” button in order to limit the output to just the desired species.

Figure 5. **Obtaining the results.** Screenshot of the different types of analysis options provided by DAVID for a given geneset. For the purpose of this tutorial, the relevant output is the “Functional Annotation Table” located at the bottom of the page.
Figure 6. Interpreting the output. Screenshot of the DAVID “Functional Annotation Table” results for a geneset. Top left: (blue bar): Gene Symbol identifier. Center (blue bar): full gene name. Labels (left): GO Term BP, GO Term CC, and GO Term MF descriptors. Note that these descriptors are clickable.

Figure 7. Keyword search. Screenshot of a keyword search of the DAVID output to identify genes that are relevant to a given biological function. In this case the search term is “neur” (highlighted in yellow), to identify genes related to neurological processes. The keyword search is done via the general search function on your computer. For larger genesets, a computer program may be used.
random list of genes for evaluation. (See also Video 2 for instruction (Delprato et al., 2019b))

**Protocol**

Step 1. Specify species: Human is the default

Step 2. Specify list length: 200-500 is a good representative number. Note that DAVID will not evaluate lists with more than 2000 genes. An error message stating this will be received.

Step 3. Select “Generate”

Step 4. Copy the gene list using the “Select All” option and paste the list directly into DAVID for evaluation as described above. Make sure to select “Official Gene Symbol” as the identifier when submitting the gene list.

**Learning assessment**

A basic entry and exit ticket method is suggested to determine what students know about genomics and genes before the lesson as well as what they have learned: main points, questions they may have and what they found most interesting. Sample questions are provided in what follows.

**Entry ticket questions and answers**

1. What is a gene?

A sequence of DNA or RNA which codes for a molecule that has a function. A gene is the basic physical and functional unit of heredity

2. What is genomics?

Study of the full set of the genes and DNA in an organism

3. How many protein coding genes does a human have?

~20,000

4. Do humans all have the same genes?

Yes, but people have different alleles. Alleles are the variation of a gene resulting from mutations. As an example consider eye color. We all have the gene for eye color but some of us have brown, blue or green eyes and there are different shades and hues within those categories.

**Exit ticket questions**

Step 1. What were the main points of the lesson?

Step 2. Do you have any questions?

Step 3. What aspect of this lesson did you find most interesting?

**Conclusions**

We describe a procedure for students to become acquainted with the Gene Ontology classification systems which is widely used in genomics and systems biology research to characterize gene function. Grouping genes with GO Terms and the DAVID database is based on a protocol that we use with our summer interns to profile gene expression data related to behavioral neuroscience studies (Crusio et al., 2017). Grouping genes in this way identifies genes that function in the same processes and also provides information about the overall distribution of a set of genes associated with a particular tissue or process. This tutorial will familiarize early stage students with a biological database and teach them how to mine and extract useful information from a sample list of genes. Entry and exit ticket questions are also included as a formative assessment strategy.

**Data availability**

**Underlying data**

All data underlying the results are available as part of the article and no additional source data are required.

**Extended data**

Extended data is available from figshare

Figshare: Extended data 1. Video 1: GeneSetProfiling Instructional video for using DAVID to obtain Gene Ontology classifiers for a sample geneset which is provided by the DAVID site https://doi.org/10.6084/m9.figshare.7649225.v1 (Delprato et al., 2019a)


**Grant information**

The author(s) declared that no grants were involved in supporting this work.

**References**


Crusio WE, Rubio C, Delprato A: Engaging high school students in neuroscience research through an e-internship program [version 2; referees: 3 approved]. F1000Res. 2017; 6: 20. Published Abstract | Publisher Full Text | Free Full Text


Reference Source
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com