UNIVERSITY OF CALIFORNIA
SANTA CRUZ

COMPUTATIONAL METHODS FOR DEDUCING BIOLOGICAL PROCESSES INVOLVED IN WOUND HEALING FROM GENE ANALYSIS

A thesis submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

SCIENTIFIC COMPUTING & APPLIED MATHEMATICS

by

Eliana Phillips

December 2021

The Thesis of Eliana Phillips is approved by:

_____
Professor Marcella Gomez, Chair

_____
Professor Peter Alvaro

_____
Professor Dongwook Lee

_____
Professor Vanessa Jonsson

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Contents

List of Tables

List of Figures

Computational methods for deducing biological processes involved in wound healing based on gene analysis

Eliana Phillips

Abstract

The Gene Ontology (GO) is a set of uniquely identified biological processes defined by a set of genes and organized hierarchically. Overrepresentation analysis is commonly used to determine the statistically significant GO terms assigned to a list of genes. However, this method has some drawbacks to identifying the most significant biological processes from a list of differentially expressed genes from microarray data. Namely, many GO terms are highly overlapping, and many GO terms are too vague or too specific to provide meaningful interpretation. In this work, I develop a pipeline to derive a shortlist of GO terms obtained from overrepresentation analysis. I do this in two steps, "representation filtering" and "similarity filtering." First, I use information theory to quantify specificity of GO terms, and define metrics to quantify representation of GO terms in the overall dataset. These metrics are used to reduce the list in the representation filtering step. Second, I obtain pairwise similarity scores of GO terms from the NaviGO, and use these scores to perform the similarity filtering step, which eliminates redundancy in the list. This pipeline is applied to overrepresentation analysis of time-series transcriptomic data in wound healing in mice and humans. By analyzing the resulting lists of GO terms at each time point measured, I show that the shortlists significantly reduce GO list size, yet provide concise descriptions of expected wound healing stages. The main takeaways and conclusions from this study include: significant overlap between inflammation and proliferation is evident; proliferation related processes are more pronounced and varied in humans than in mice; and that inflammation is relatively consistent across datasets, but may appear to be prolonged depending on the thresholds set for differential expression of genes. This method provides a tool that allows data from transcriptomic studies to be used in translational research. In future work, the tool may used for other experiments involving time-series transcriptomic data.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Marcella Gomez, for her support and seeing this project through over many months. The completion of this thesis would not have been possible without her expertise, guidance, and mentorship. I'd also like to extend my sincere thanks to my other reading committee members Peter Alvaro, Dongwook Lee, and Vanessa Jonsson for their support and feedback on this paper. Many thanks also to Ksenia Zlobina for her help in obtaining the data that I used in this project as well as her valuable input. I am also very grateful to my Mom and Dad for supporting me and listening to me practice my defense over and over, and my friend Rabeya for encouraging me to find joy in scientific research. Finally, a special thanks to my friend Shea for inspiring me to investigate the probability of co-occurrence of GO terms through connect-a-cave.

1 Introduction & motivation

1.1 Gene analysis

The main principle or "central dogma" of molecular biology is the key to understanding how all living organisms are dependent on their genetic code. This principle states that DNA produces messenger RNA, which in turn codes for the amino acids that bond together to form proteins, the macromolecules that perform nearly all biological processes in all organisms. Understanding the genes and pathways driving gene expression helps us better understand complex biological processes. Identifying prominent biological processes from differentially expressed genes can help scientists define stages and transitions in complex macro level processes more rigorously. For example, the process of wound healing occurs in four main stages with many overlapping sub-processes in each stage; in this work I use gene analysis techniques and develop a computational tool to map biological processes to stages of wound healing.

Gene analysis methods may be helpful in overcoming the barriers of using transcriptomic data for translational research and medical discovery. Translational research is motivated by clinically relevant problems and consists of developing medical diagnostics, procedures, and technologies. Consequently, this field must be highly collaborative and interdisciplinary in order to bridge the gap between theory and application. This gap is often called the "Valley of Death", as it is where early biological discoveries go to die before they can be applied to medical practice.

Transcriptomics refers to studies involving the set of all RNA transcripts in an individual or population of cells. Most studies involving transcriptomic data are constrained to identifying differentially expressed genes, and do not provide functional interpretation of how gene expression changes drive complex biological processes. There is an unmet need for introducing new tools to draw conclusions to assist in translational science.

1.2 Gene ontology

To help bridge this gap between transcriptomic data and functional knowledge of biology, the Gene Ontology (GO) database is commonly used. GO is actually composed of three ontologies, which are mostly disjoint from one another. These are biological process, defined as higher level processes accomplished by multiple molecular activities working together; molecular function, or molecular level activities performed by gene products; and cellular component, which are the locations or cellular structures in which a function takes place. Each aspect is comprised of a set of terms with relations operating between them, and each term is annotated to one or more genes in a given species. The database is dynamically updated to reflect current knowledge and new annotations are added frequently. In this work, I will be focusing on the biological process ontology only for simplicity and relevancy.

GO is organized as a directed acyclic graph (DAG), where the nodes are GO terms and the edges are relations between them. The direction of the edges determines whether, in a given pair of nodes A and B, A is a parent of B or vice versa. A DAG is a more general mathematical structure than a tree, in which each node can only have one parent. In a DAG, any node can have any number of parents and children as long as it contains no cycles. If a term is a parent of another term, then the function it represents is higher level, or more general, than its child. The root of the graph, or the node with no parents, is the highest level GO term, simply "biological process." The nodes with no children are often called the leaves of the graph; these nodes represent the most specific or low level processes. I will introduce specific ways of measuring the specificity of nodes in the GO in section 2.

1.3 Overrepresentation analysis

The online PANTHER database tool uses the GO annotations to perform overrepresentation analysis on lists of genes. Overrepresentation analysis is a statistical method that identifies a list of GO terms for a given gene list and determines whether each GO term is significantly over or underrepresented in the list of genes. This is also sometimes called enrichment analysis. This tool

is extremely useful for extracting significant biological information from a list of genes. For example, we may use it to determine what types of proteins were produced during specific times over the course of a laboratory experiment, based on which genes were expressed.

To use the tool, we input a list of genes that we wish to analyze, ideally sharing common characteristics such as the top differentially expressed genes at a given time point to obtain more refined results. We also input a reference list of genes; in our case this is the list of all the genes measured from the microarray experiment that our input list was taken from. The default reference list in PANTHER is the entire genome for the given species, however, results of overrepresentation analysis are much more accurate when the reference list is set according to the microarray chip used. Otherwise, the resulting list of GO terms will contain spurious results.

The outputs of overrepresentation analysis in PANTHER consists of statistical information as well as biological. Each result includes, but is not limited to, the following fields:

- The GO term itself, consisting of a unique numerical label and the biological process it represents.
- The list of genes in the inputted list that are annotated to that GO term.
- Level in ontology, indicating how far the GO term is from the root of the graph (biological process). This gives us an idea of how specific that GO term is in the overall DAG. The lowest level terms, level 0, are farthest from the root, while the highest level terms are closest to it. This is somewhat counterintuitive; to make matters more complex, the more general nodes closer to the root may appear to have different levels depending on how many ancestors are present along different paths of the DAG. We will expand upon this issue in section 2.2.
- A + or -, indicating if the term is over or under-represented, respectively. This allows us to easily filter out the underrepresented terms for analysis.

- The p-value of the term, which is the likelihood that the mapping of that GO term to the sub-list of genes occurs purely by chance, as opposed to carrying some biological significance. Specifically, it is the probability of seeing at least x number of genes out of the total n genes in the list annotated to a particualar GO term, given the proportion of genes in the whole genome that are annotated to that GO term. All terms with $p > 0.05$ are omitted in the output, thus what remains is statistically significant terms only.

- The number of genes in the input list mapped to that GO term.

- The number of genes in the reference list (the microarray data) mapped to that GO term.

The purpose of this analysis is to learn which GO terms are most or least significant in the overall functions of a set of genes with similar dynamic characteristics. In the case of our data, these characteristics are peak time of expression over the course of the experiment and intensity of expression. Since many GO terms are present in a given list of genes, we can use this tool as a mathematical reference for their relative significance and apply quantitative reasoning to arrive at a concise yet complete list. Overall, it helps in the study of translating transcriptomic data into useful biological information.

1.4 Previous work in GO analysis

The GO itself, which can be found at geneontology.org, contains all current knowledge of GO terms and annotations and is updated regularly by experts in the field. It is a computational representation of our current scientific knowledge about the functions of genes from many different organisms. Users can perform overrepresentation analysis through PANTHER on the GO website, as well as look up any gene or GO term.

The paper "Use and Misuse of the Gene Ontology Annotations" by Rhee et al, explains some of the drawbacks of the current approach to gene ontology. I will highlight a few of the main drawbacks relevant to this work. First, some information in GO may be imprecise, as some annotations are made at

very high level GO terms, which limits their usefulness in drawing meaningful conclusions from the data. There is also always the possibility that some annotations are not correct and not able to be corrected by human experts or automated tools. Some genes are also involved in several different biological processes, but the tool weights all processes equally and cannot single out the most relevant process with respect to the data using the context of the other genes in the list, which can lead to extraneous results. Additionally, existing approaches are decoupled from gene expression data obtained from transcriptomic studies. Different genes may be regulated or expressed to different extents, however, GO does not provide information on how the amount of gene regulation correlates with the intensity or relevancy of the corresponding biological process. Finally, since some biological processes are more well known than others, they have more annotations to genes and thus are more likely to appear significant than others with less annotations, even if this is not the case.

Other previous work related to GO includes the development of databases designed to obtain a greater understanding of GO, namely, the GO Partition Database (GO PaD) and the NaviGO database. The GO PaD provides a reorganization of the GO by using information theory to partition the GO into subsets based on their information content in bits. Applications of this study to our work is expanded upon in section 2.2. NaviGO introduces similarity scores that allow for pairwise comparison between GO terms, and the database itself allows us to enter a list of GO terms and obtain these scores to use for analysis. Section 2.3 contains further explanation on how this tool is applied to our work.

2 Technical Background

2.1 Selecting gene lists for analysis

The data used in this study was collected using publicly available DNA microarrays; these consist of small chips containing thousands of DNA sequences from a particular organism. These datasets can be found here. Microarrays

are used to study genetics, specifically for transcriptomics in order to determine which genes are expressed in certain cells and tissues. In this report, I will explore the application of time-series transcriptomic data analysis for the study of wound healing. From microarray data, I select the most highly differentially expressed genes for my analysis.

A gene is considered differentially expressed if a difference observed in read counts or expression levels between two experimental conditions is statistically significant. The expression level of a gene is proportional to the amount of protein it produces. Studying the genes that are differentially expressed over the course of a given experiment, such as during different intervals of time, allows us to understand the underlying biological processes involved in that experiment. Differentially expressed genes may also be called upregulated or downregulated, depending on if its change in expression is positive or negative. Upregulation is the process by which a cell increases the quantity of a gene product/cellular component, such as RNA or protein, in response to an external stimulus. The decrease of such components is downregulation. Consequently, upregulated genes in an experiment at a given time point are the ones which produce high amounts of gene product and are highly differentially expressed. For example, genes that code for the production of proteins crucial for the body's immune response will be highly upregulated during the inflammation stage of wound healing, and less upregulated as the process continues to later stages.

Since the microarrays we draw from contain tens of thousands of genes, many are not significant or highly expressed enough to consider for analysis. We can extract the most highly differentially expressed genes from the list by computing the maximal fold change, or difference in gene expression intensity over the given time points during an experiment. We then filter the genes according to their maximum fold change in intensity relative to the start of the experiment with respect to a given threshold as detailed by Zlobina et al. The resulting gene lists after filtering may then be used for overrepresentation analysis in order to obtain the significant GO terms associated with them.

2.2 Information content in bits as a metric for GO term specificity

The GO Partition Database and corresponding paper, by Alterovitz et al, proposes using information content of a GO term in bits as an alternative way of quantifying specificity of GO terms rather than using the level in ontology as described in section 1.3.

Testing biological hypotheses typically assumes comparable specificity or levels in the DAG. However, DAG levels are not good indicators of specificity, as they may cause misleading results or miss important biological discoveries. One such reason for this is that the level in the ontology is not a set quantity. Since GO is a DAG, following different paths of ancestry may suggest a single term has multiple levels. PANTHER uses the path followed through the graph to determine the level of a GO term in the results for a particular gene list. This means that the same GO term may appear to have more than one distinct level. This makes it difficult to assess how specific or general that GO term is without the context of the path followed, which PANTHER is not able to determine. This can lead to inconsistent discoveries. Additionally, within any single level, there is a lot of variation of specificity; for example, level 3 in one set of GO terms may not be comparable to level 3 in another set. Again, this is a result of the DAG structure and the fact that some pathways in the DAG may have more ancestors or descendants than others. The GO subgraph shown in figure 1 below illustrates these issues. This motivates the use of information theory to quantify specificity of GO terms.

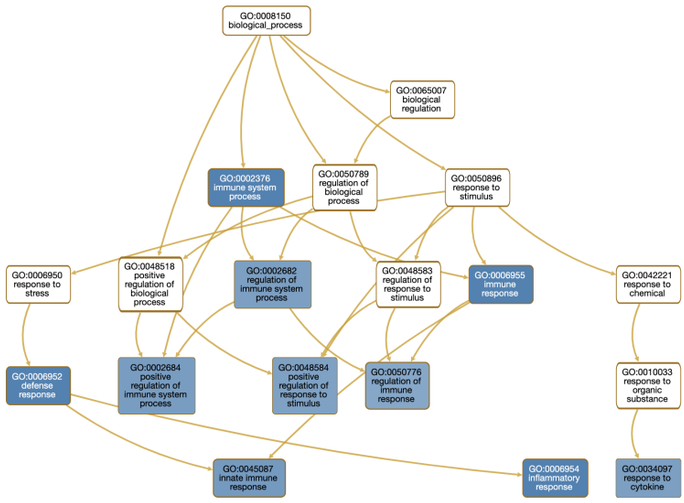


Figure 1: Example DAG showing subgraph of the GO for top 100 DE genes day 1 post-wounding of mouse skin.

2.2.1 Overview of information theory

Information theory, the study of the quantification, storage, and communication of digital information, is commonly used in mathematics, computer science, statistics, and other fields. The fundamental unit of information is the bit (a contraction of "binary digit"), which represents a logical state of a system with the value of either 0 or 1. Central to the study of information theory is entropy, a quantification of the amount of uncertainty involved in a system or a random variable. The more possible discrete states a system has, the higher its entropy, and the lower the probability of choosing a given state at random. We will see that high entropy systems contain high information content. Thus, we can understand the bit as the quantity of information stored in a binary random variable that has an equal probability of taking the value 0 or 1. Eight bits is equal to a byte, which constitutes the base unit typically used to denote storage in computing as with megabytes, gigabytes, and so forth. Any data, digital or not, can be encoded using bits, providing a universal unit by which to quantify more abstract information that is easily translatable to computational studies.

Claude Shannon, one of the founders of information theory, formalized that information that can be stored in a system is proportional to $\log_b N$, where N is the number of states in a system and b is the unit used to measure informa-

tion. This is key to understanding how the information of any system relates to its entropy, and the use of information theory as a standard measure to be applied to more abstract data.

2.2.2 Application to GO

Information content in bits can serve as a proxy for specificity of GO terms, with high level or less specific terms containing few bits of information, while low level or more specific terms contain many bits. Alterovitz et al illustrates the use of this information based framework to partition GO into sets identified by uniform information content to create a new database for organizing GO data, and shows that information is much more evenly distributed across a set of GO terms containing information content in a given interval of bits than a set of GO terms said to have the same level in the DAG. While the database itself is not relevant to this work, the concept of quantifying GO terms by their information in bits is a key part of this study.

Recall that the information stored in a piece of data is dependent on the entropy of the system containing that data. In the case of GO, the system is the entire ontology of terms and the data refers to a single GO term. Let a GO term be represented as V_n, the gene set annotated by V_n be $k(V_n)$, and j be the total number of GO terms in the ontology. We can then define the probability $p(V_n)$ of randomly selecting a gene from the entire microarray that is annotated to V_n as follows. In the below equation, the numerator is the number of genes in the microarray annotated to V_n, and the denominator is the total number of genes in the microarray,

$$p(V_n) = \frac{k(V_n)}{\bigcup_{m=1}^{j} k(V_m)}. \tag{1}$$

The information content in bits $I(V_n)$ of GO term V_n is then computed as follows. This quantity is also called Shannon information, after Shannon's formalization of information stored in a system.

$$I(V_n) = -\log_2 p(V_n) \tag{2}$$

The logarithmic relationship between probability and information content means that an increase in one bit of information corresponds to a 2-fold increase in specificity. The probability of selecting a GO term with 0 bits of information is 1, for 1 bit of information, the probability is 1/2, and so on. We can also say that a GO term with $I(V_n) = n$ has a probability of $1/2^n$ of being randomly selected.

Using the GO data provided by PANTHER, we can easily compute the probability of selection and information content of any given GO term. For a given microarray, the number of genes in the microarray is constant, typically around 40,000. The number of genes in the microarray annotated to a GO term is given by the "number_in_reference" field in the GO output. During the data parsing phase of the pipeline that will be explained in detail in section 3 of this report, we assign information content to each GO term and use it to quantify the specificity of each term as compared with its representation in the dataset.

2.3 Similarity and comparison of GO terms

The NaviGO database provides many advantages to further understanding the relationships between GO terms. Built into the tool is an interactive rendering of the GO DAG that provides an intuitive understanding of similarity among them. The novel aspect of this tool is six GO similarity scoring schemes that reflect different types of pairwise relationships between GO terms. These relationships include, but are not limited to, the topological structure of GO, protein-protein interactions between gene products annotated to GO terms, contextual association, which provides a metric of how often the two GO terms appear together, and annotation frequency. The quantitative analysis of GO term distance and functional similarities provided by this database is invaluable to this work.

The six scores consist of three semantic similarity scores, which quantify the closeness of the two GO terms in the GO DAG, and three functional similarity scores, which quantify how likely two terms are to interact or co occur with one another. Functional similarity can also often be inferred from the semantic

similarity scores as well, for example, if the score is low, this indicates that the terms are far away from each other in the DAG, and thus, may perform very different functions. On the other hand, the interaction association score (IAS), one of the functional similarity scores represents the degree to which the proteins annotated to a pair of GO terms interact, which may be very high even when the semantic score is low as the IAS can identify related terms across GO categories.

We select two of these six scores to use for our analysis, one semantic similarity score and one functional similarity score. The methods and results section of this paper will expand on the use of these scores to further our research.

For semantic similarity, we use the relevance semantic similarity score (RSS), which calculates the information content in bits of the lowest common ancestor of two terms in the DAG. If the lowest common ancestor of the two terms has a low information content, then the two terms are likely far away and their RSS score is low; if it has high information content, then the two terms are likely close together and the score is high. The score is calculated as follows.

$$sim_{Rel}(c_1, c_2) = \max_{c \in S(c_1,c_2)} \left(\frac{2 \log p(c)}{\log p(c_1) + \log p(c_2)} \cdot (1 - p(c)) \right) \tag{3}$$

Recalling how information content is computed from probability, $p(c)$ represents the probability of identifying GO term c. The first term computes the relative depth of the common ancestor c to the depth of the two terms c_1 and c_2 while the second term represents how rare it is to identify the common ancestor c by chance.

We will use the co-occurrence association score (CAS) for the functional similarity score. CAS was designed to quantify the frequency of co-occurrences of two GO terms in a single gene annotation relative to random chance and is computed as follows.

$$CAS(i, j) = \frac{\frac{C(i,j)}{\sum_{i,j} C(i,j)}}{\left(\frac{C(i)}{\sum_k C(k)}\right)\left(\frac{C(j)}{\sum_k C(k)}\right)} \tag{4}$$

Here, $C(i, j)$ is the number of sequences of gene annotations in the Gene Ontology Annotation database that contain both the GO terms i and j. $C(k)$ is the total number of sequences in the GOA annotated with GO term k for $k = i, j$. In short, the CAS quantifies how often two GO terms i and j co-annotate sequences relative to random chance. CAS also includes GO hierarchy information in scoring the term pairs.

3 Methods

3.1 Gene ontology significant terms pipeline

The main novel contribution of this work is a pipeline, created as a tool in Python, that takes in lists of differentially expressed genes to be subjected to overrepresentation analysis, and narrows down the resulting long list of GO terms into a short list of the most significant GO terms representing that list of genes.

3.1.1 Motivation for filtering GO terms

Many GO terms that result from overrepresentation analysis are much too general to provide useful information about the biological function they code for, such as the root node "biological process", or other high-level terms such as "metabolic process" and "immune process". On the other hand, some GO terms are too highly specific, meaning that very few genes in our input list are annotated to that term. In this case, we cannot assert with confidence that the associated biological process is in fact significant. Having such a small sample size also increases the risk that the gene in question may have resulted from experimental noise. Additionally, many GO terms are redundant, and all of these redundant terms will show up when performing overrepresentation analysis on a given gene list. This redundancy may be caused by genes in the data that are annotated to multiple GO terms, or two or more GO terms closely related in the ontology (such as parent-child relationships) showing up in the result list. The GO significant terms pipeline consists of two main steps, representation filtering and similarity filtering, to narrow down the list of GO terms resulting from overrepresentation analysis.

3.1.2 Representation filtering

First, data from GO is used to filter out terms that do not meet our prescribed criteria regarding specificity and representation of the term in the dataset. Let us denote the number of genes in the reference list, or microarray, annotated to a particular term as N_{GO}, and the number of genes in our input gene list as N_{input}. The intersection r of these two quantities, as shown in the below Venn diagram, represents the number of genes in our input list annotated to that GO term.

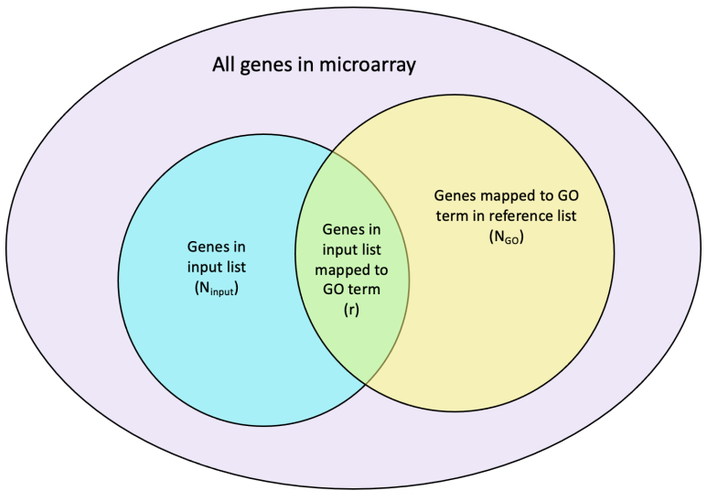


Figure 2: Venn diagram detailing representation metrics of GO terms.

We define two metrics to quantify how represented a GO term is in the dataset. Let representation amount of a term be $R_a = r/N_{GO}$, and normalized representation amount of a term be $R_n = r/N_{input}$.

The two histograms below show the distributions of the representation amount and normalized representation amount of GO terms for successive intervals of information content for day 1 mouse skin wound top 100 differentially expressed genes. Further explanation of how these histograms are generated is given in section 3.2; we show them here to illustrate the relationships between the representation metrics and information content. Distributions for other datasets follow similar patterns.

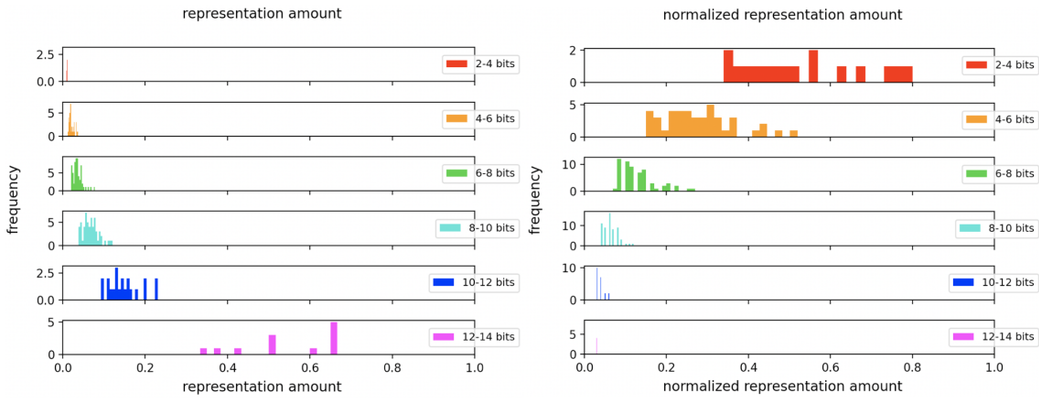


Figure 3: Example histograms showing representation metrics for different information content intervals.

These histograms indicate that the two quantities R_a and R_n appear to be inversely related. While R_a favors terms with very high information content, R_n favors terms with very low information content. We postulate that desired GO terms for analysis will be those on the more specific, or higher information end, however, recall that if a GO term has a high R_a and low R_n, this means that very few genes in our input list are annotated to that term. Therefore, we can't be too confident that the associated biological process is in fact significant. This motivates us to consider both metrics in our selection process. We want to balance choosing highly specific terms with how frequently they appear in our input list of differentially expressed genes and thus how represented they are in our list.

To approach this more quantitatively, we define a new metric, the "representation score" R_s of a GO term, as a quadratic equation that is a function of both R_a and R_n. We take the product of the two original metrics, each shifted by a user-determined bias constant, and scale it by 100 to get the following equation:

$$R_s = 100 \cdot (R_a - b_1) \cdot (R_n - b_2), \tag{5}$$

where b_1 and b_2 are biases chosen to balance how much R_s should depend on each quantity. The purpose of using biases is two-fold. Firstly, we choose the biases such that, when we plot representation score versus information content in bits over all GO terms, we obtain a concave down parabola with a clear

maximum, in order to give us an optimization problem. A monotonically increasing or decreasing function does not yield such an optimization. Second, the ranges of values that R_a and R_n can take is different. Using biases allows us to find a proper threshold for the metrics to balance this discrepancy in ranges. It is important to note that, while we don't want to exclude highly specific terms from our analysis, this may be a trade-off of considering more highly represented GO terms.

Since we want to choose biases that yield a concave down quadratic with a maximum, we must empirically determine where that maximum should lie. The example scatter plots in figure 4 below show the representation amount, normalized representation amount, and representation score of GO terms for day 1 mouse skin wound top 100 differentially expressed genes as a function of information content in bits. In observing these plots, we confirm the inverse relationship between the representation amount and normalized representation amount. We choose the maximum of our quadratic to lie approximately at the inflection point where R_a begins to increase and R_n begins to flatten out with respect to bits. We may need to adjust the biases chosen for different datasets. This occurs when considering larger gene lists for which the resulting GO terms from overrepresentation analysis skew on the lower information side. For the data shown, which is again the top 100 differentially expressed genes day 1 post-wounding of mouse skin, this happens around 10 bits, which we can see circled in red.

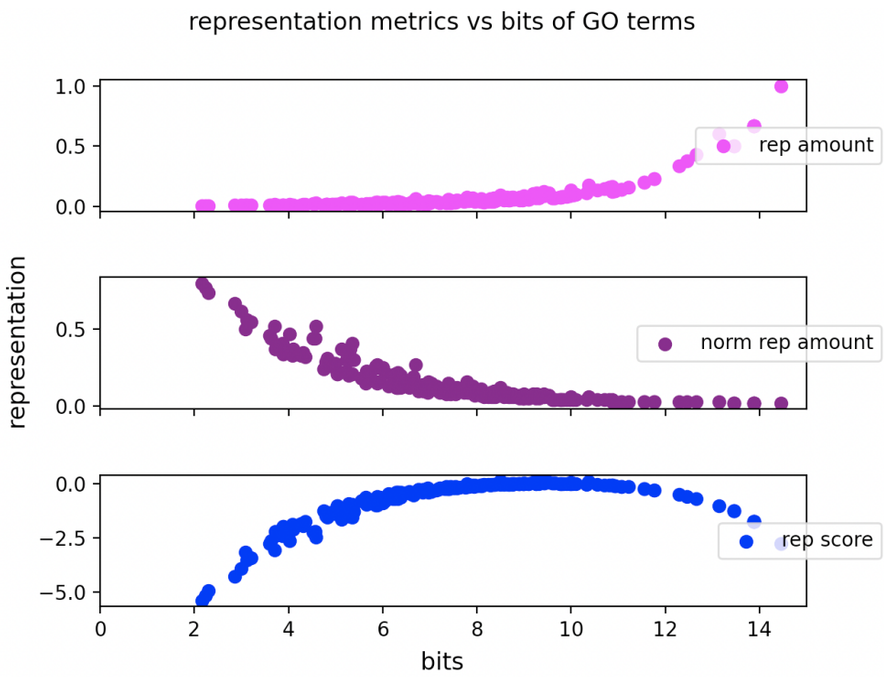


Figure 4: Example scatter plots showing representation metrics and representation score vs information content in bits.

Finally, we take the GO terms whose representation scores are in the top 10% of all scores, signifying the maximum of the quadratic shown above, for the given set to move on to the second step, and discard the rest.

3.1.3 Similarity filtering

Similarity filtering consists of using the NaviGO database to perform pairwise similarity analysis and scoring on the resulting list of GO terms to filter out sufficiently similar terms. This allows us to mitigate the issue of redundancy of GO terms in overrepresentation analysis. NaviGO gives semantic and functional similarity scores, and we choose the relevance semantic similarity score (RSS) and co-occurrence association score (CAS) to use for filtering, as explained in section 2.3. In filtering out GO terms that are similar to one another, we consider both of these scores, with slightly more weight on the semantic score. This is due to the relative lack of information in the database for the CAS for many pairs, and because high semantic similarity scores can identify parent-child pairs in the GO set, which are a common source of redundancy in overrepresentation analysis.

We consider a pair to be similar if one of the following criteria is true. These

criteria are informal heuristics we tested over several iterations of the tool's development and subsequently defined to select sufficiently similar pairs. A pair of GO terms is considered similar if either the RSS is greater than 0.9, where the range for RSS is 0 to 1, or the RSS is in the top 10% and the CAS is in the top 25% of all scores in the group.

Similar pairs are placed in similarity groups based on mutual similarity with other terms. The graph below shown in figure 5 is an example of a similarity network graph of GO terms for day 1 mouse skin wound top 100 differentially expressed genes. There are two connected components, each representing a separate similarity group. There component to the left contains only two terms, allowing us to choose one to represent the group. The larger component to the right shows a more complicated structure from which we mathematically discern the best term to represent the group.

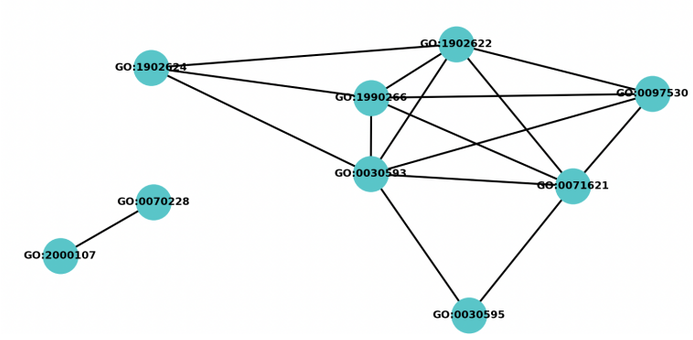


Figure 5: Example network graph showing similarity groups of GO terms.

To determine which terms appear in the shortlist, we use another heuristic. This criteria is as follows. All GO terms that appear in none of the similarity groups are added to the shortlist. This allows us to avoid eliminating any possibly significant biological processes that do not contain any redundancy in the set, which is always possible with GO analysis. For each similarity group, one GO term is chosen to represent the group and the rest are discarded. For each similarity group, the GO term with the highest degree, or highest number of connections to other terms, is chosen. The node circled in red in the graph shown is the one with highest degree for that component of the graph. If there is more than one node with maximum degree, or if the component is a complete graph, then we select the max-degree term with the highest

information content.

3.2 Generating visualizations

As illustrated in the examples shown in above sections, visualizations help us to greater understand the abstract concepts we utilize in this paper and apply to practical work. We use statistical data involving the relationships between representation of GO terms and information content to create histograms and scatter plots, and similarity data from NaviGO to create similarity network graphs. The histograms, scatter plots, and similarity graphs for all other time points for top 100 and 1000 mouse genes and top 100 human genes can be found in the appendix at the end of this paper.

Histograms allow us to visualize the relationships between information content in bits of sets of GO terms and the values R_a and R_n we defined to quantify representation in the dataset of those terms. During data parsing, when we compute the information content in bits of each term, we partition the set of GO terms into the intervals 0-2, 2-4, ..., 12-14 bits, where nearly all observed GO terms contain less than 14 bits of information.

Additionally, we employ the use of scatter plots to see how the representation amount, normalized representation amount, and representation score behave as functions of information content in bits. There are two main motivations for generating this visualization. First, it is easier to observe the overall trends in R_a and R_n vs bits as a scatter plot than as a histogram. This confirms our earlier observation that the two quantities seem to be inversely related. Secondly, and most importantly, comparing the representation score R_s to each of its constituents allows us to qualitatively observe how much R_s resembles R_a and R_n, and adjust the biases if need be to balance the visibility of each. Since R_s is a quadratic model, we also can adjust the biases to construct a concave down function with a clear maximum in the desired bit range, ensuring that when we filter out the top 10% most highly represented GO terms, they will have the specificity we require.

We also generate graphs to show the similarity between terms identified by NaviGO and used for similarity filtering. Each GO term is represented by a node and edges between them represent similarity by our criteria. This allows us to see which GO terms are similar and what the structure of the similarity groups looks like. Some groups are complete subgraphs, meaning all terms are similar to each other. This means we can choose one term arbitrarily to represent the group, so we choose the terms with the most information out of the group. Otherwise, if one term has more connections than the rest of the terms in the group, we choose that term to represent the group.

3.3 Python tool

The Python tool used for this analysis consists of several codes to perform data parsing, figure generating, representation filtering, and similarity filtering. We use the tool as follows. First, we obtain the lists of the most highly differentially expressed (DE) genes for a given organism at a given peak time. We then enter each list into PANTHER with the corresponding reference list of the microarray data that those DE genes came from. The resulting GO terms in the form of the JSON output of PANTHER are entered into step 1 of the pipeline by entering the JSON file and the "rep" keyword and running the script. This file is parsed to obtain the relevant information used for representation filtering. The script will output a spreadsheet containing the filtered GO terms and their corresponding data. We enter these GO terms into the NaviGO database to obtain the similarity data and download this result as another spreadsheet, which is then fed into step 2 of the pipeline along with the "sim" keyword. The script finally outputs the shortlist as another excel spreadsheet detailing the results of the shortlist and corresponding data, consisting of each GO term and its information content, representation amount, normalized representation amount, and representation score. All codes and supporting materials and files are available on GitHub at this repository.

4 Results

4.1 Application to wound healing

We use the GO shortlisting pipeline described in the previous section to analyze time-series transcriptomic data of mouse and human wound healing. We produce shortlists of GO terms for both species at each time point recorded in the microarray experiments using lists of the most highly differentially expressed genes at each of these time points. The goal is to map wound healing stages to dynamic processes based on the genes expressed. Identifying prominent biological processes involved from differentially expressed genes may help us define stages and transitions more rigorously.

The main goal of the work in the Gomez lab and overall collaboration is to create a bioelectronic bandage that stimulates and accelerates wound healing. In order to do this, collaborators are working on a predictive model using images and gene expression data to determine wound healing stage progression. Thus, mapping biological processes to wound healing stages to augment our predictive model is a main application of this work.

4.1.1 Overview of wound healing

The wound healing process consists of four main stages. The first stage is hemostasis, which lasts for a few hours immediately post-wounding. During this time, blood vessels constrict to reduce blood flow to the wound and prevent the body from bleeding out and platelets stick together to repair the blood vessel. Then, coagulation occurs and a clot composed of blood and fibrin polymers forms over the wound. The next stage, inflammation, begins within the first day of wounding and can last anywhere from several days to several weeks. The injured blood vessels cause localized swelling and immune cells such as T cells, leukocytes, and neutrophils migrate to the site of the wound to defend the tissue from infection. Damaged cells and pathogens are also removed. Swelling, heat and pain are normal during this stage, paving the way for the wound to be rebuilt during the proliferative phase. During this stage, the wound contracts, new tissues are rebuilt with collagen and extra-

cellular matrix proteins, and new blood vessels are constructed to replace the damaged ones. Tissue repair continues with chemotaxis of immune cells. [verify this] At the end of proliferation, the wound site epithelializes, resurfacing a new layer of skin. The final and longest stage is maturation, or remodeling. This begins a few weeks after the injury and may continue for a year. Maturation occurs when the collagen at the wound site is remodeled and and the wound is able to fully close. This remodeling allows the collagen fibers to organize and cross-link, strengthening the skin of the wound. Programmed cell death of repair cells that are no longer needed also occurs during this time.

There is a lot of overlap between these stages. For example, some processes involved in inflammation may still be occurring as the proliferative phase starts, such as immune cell migration. For this reason, it is often difficult to say when one stage ends and the next begins. Intermediate stages may also be defined to address this difficulty. Identifying the genes and pathways of gene expression involved in each stage of the wound healing process helps us to better understand each stage, overlap between stages, and how the many different molecular-level processes work together to achieve the desired results. Lists of the most highly differentially expressed genes from DNA microarrays profiling wound healing were used for this analysis.

4.1.2 Datasets used

The National Center for Biotechnology Information's transcriptional profiling of the wound healing process in mice and humans by way of Affymetrix GeneChip microarrays is available online for public use. Three microarray experiment datasets were used in this study, one mouse and two human. Series GSE23006 examined mouse skin and tongue wound tissue samples from days 0 to 10 after wounding, spanning all stages of the wound healing process; we study the skin wound data only. Samples were taken 6 hours, 12 hours, 1 day, 3 days, 5 days, 7 days, and 10 days post wounding. Series GSE28914 examined human skin wound tissue samples from 8 patients 3 and 7 days after wounding, covering the inflammation and beginning proliferation stages of wound healing only. Series GSE50425 also studied human skin wound tissue, focusing on the

maturation (final) stage of wound healing. Samples were taken from 4 patients 14 and 21 days post wounding.

Since the microarrays contain large numbers of genes, we filter the most highly differentially expressed genes to consider for analysis. This is done by calculating the fold-change in expression of each gene. The top 100 most differentially expressed genes for each recorded time point in both mouse and human microarrays were used for the GO shortlisting pipeline. We also performed the analysis on the top 1000 most differentially expressed genes for the mouse data for comparison.

4.2 Shortlists, top 100 differentially expressed genes

In the following section, we present the shortlists of the most significant GO terms for the mouse and human datasets obtained from lists of the top 100 most differentially expressed genes at each recorded time point.

4.2.1 Mouse data

For the following dataset, we set biases $b_1 = 0.08$ and $b_2 = 0.05$ for the representation filtering stage.

Table 1: GO shortlists for each time point in top 100 most differentially expressed mouse genes.

Peak time	GO ID	Biological Process
6 hours	GO:0030593	neutrophil chemotaxis
	GO:0007159	leukocyte cell-cell adhesion
	GO:0071674	mononuclear cell migration
	GO:0030593	regulation of interleukin-8 production
	GO:2000379	positive regulation of reactive oxygen species metabolic process
	GO:0072676	lymphocyte migration
12 hours	GO:0030593	neutrophil chemotaxis
	GO:0007159	leukocyte cell-cell adhesion
	GO:0072676	lymphocyte migration
	GO:0032757	positive regulation of interleukin-8 production
	GO:0032653	regulation of interleukin-10 production
	GO:0070098	chemokine-mediated signaling pathway
1 day	GO:0002526	acute inflammatory response
	GO:0007159	leukocyte cell-cell adhesion
	GO:0030593	neutrophil chemotaxis
	GO:0050766	positive regulation of phagocytosis
	GO:0045071	negative regulation of viral genome replication
	GO:0042116	macrophage activation
	GO:0032757	regulation of interleukin-10 protein
	GO:0032757	positive regulation of interleukin-8 protein
	GO:2000107	negative regulation of leukocyte apoptotic process
	GO:0002224	toll-like receptor signaling pathway
	GO:0072676	lymphocyte migration
3 days	GO:0002675	positive regulation of acute inflammatory response
	GO:0071677	positive regulation of mononuclear cell migration
	GO:0030593	neutrophil chemotaxis
	GO:0050766	positive regulation of phagocytosis
5 days	GO:0030199	collagen fibril organization
7 days	GO:0030199	collagen fibril organization
10 days	GO:0030199	collagen fibril organization
	GO:0045766	positive regulation of angiogenesis
	GO:1904018	positive regulation of vasculature development

4.2.2 Human data

For the following dataset, we set biases $b_1 = 0.15$ and $b_2 = 0.05$ for the representation filtering stage.

Table 2: GO shortlists for each time point in top 100 most differentially expressed human genes.

Peak time	GO ID	Biological Process
3 days	GO:0002286	T cell activation involved in immune response
	GO:0002828	regulation of type 2 immune response
	GO:0030593	neutrophil chemotaxis
	GO:0070098	chemokine-mediated signaling pathway
	GO:0032673	regulation of interleukin-4 production
	GO:0002675	positive regulation of acute inflammatory response
	GO:2000108	positive regulation of leukocyte apoptotic process
	GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II
	GO:1901889	negative regulation of cell junction assembly
	GO:0043299	leukocyte degranulation
	GO:0019835	cytolysis
	GO:0045429	positive regulation of nitric oxide biosynthetic process
	GO:0080164	regulation of nitric oxide metabolic process
	GO:0022617	extracellular matrix disassembly
	GO:0034605	cellular response to heat
	GO:0052372	modulation by symbiont of entry into host
	GO:0060760	positive regulation of response to cytokine stimulus
	GO:0032663	regulation of interleukin-2 production
7 days	GO:0048247	lymphocyte chemotaxis
	GO:0150077	regulation of neuroinflammatory response
	GO:0030593	neutrophil chemotaxis
	GO:0030574	collagen catabolic process
	GO:0010575	positive regulation of vascular endothelial growth factor production
	GO:0031424	keratinization
	GO:2000403	positive regulation of lymphocyte migration
	GO:0006953	acute-phase response
14 days	GO:0034110	regulation of homotypic cell-cell adhesion
	GO:0030574	collagen catabolic process
	GO:0090288	negative regulation of cellular response to growth factor stimulus
	GO:0060412	ventricular septum morphogenesis
	GO:0010718	positive regulation of epithelial to mesenchymal transition
	GO:0014911	positive regulation of smooth muscle cell migration
	GO:0003179	heart valve morphogenesis
	GO:0061035	regulation of cartilage development
21 days	GO:0002690	positive regulation of leukocyte chemotaxis
	GO:0022617	extracellular matrix disassembly
	GO:0045778	positive regulation of ossification
	GO:0060688	regulation of morphogenesis of a branching structure
	GO:0001960	negative regulation of cytokine-mediated signaling pathway
	GO:0032963	collagen metabolic process
	GO:0032835	glomerulus development
	GO:0045071	negative regulation of viral genome replication
	GO:0032729	positive regulation of interferon-gamma production
	GO:0030510	regulation of BMP signaling pathway
	GO:0048771	tissue remodeling
	GO:0003279	cardiac septum development

4.3 Shortlists, top 1000 differentially expressed genes

To test the GO shortlisting method's efficacy in taking in different gene list sizes, we additionally enter the top 1000 most differentially expressed genes at each time point for the mouse data only. When we enter a list of 1000 genes into PANTHER for overrepresentation analysis, the most significant GO terms that come out of the algorithm are much more general than those we obtained from entering a list of 100 genes. This is due to PANTHER's criteria for selecting overrepresented GO terms for a gene list. When taking in a much larger list of genes, there is a larger probability that a highly informative GO term is annotated to a gene or sub-list of genes by random chance. In other words, it is more likely that highly informative GO terms will have larger p-values and thus not appear in the GO overrepresentation results. While plenty of specific GO terms do still appear in the results, their R_a and R_n values might be lower regardless of their significance.

To attempt to mitigate this issue, we adjust the biases in computing the representation score. Recall from section 3.1.2 that we choose biases that yield a concave down quadratic with a clear maximum, which lies approximately at the inflection point where R_a begins to increase and R_n begins to flatten out with respect to bits. Since the ranges of these metrics change with respect to gene list size, we again find biases that balance this discrepancy.

While changing the biases still may not yield the same results from the top 1000 genes as with the top 100 genes, we may obtain results that are more similar or at least more easily comparable. Again, the process of selecting the biases is done heuristically and tested for relative efficacy rather than objective truth. We then analyze the results to ensure that the biological processes represented align with the expected functions that occur at the stage(s) most likely mapped to that particular time point. This ensures relative consistency of the algorithm across different gene list sizes, as long as the biases are shifted accordingly. For this dataset, we set biases $b_1 = 0.6$ and $b_2 = 0.06$.

Table 3: GO shortlists for each time point in top 1000 most differentially expressed mouse genes.

Peak time	GO ID	Biological Process
6 hours	GO:0030490	maturation of SSU-rRNA
	GO:0090501	RNA phosphodiester bond hydrolysis
	GO:0030216	keratinocyte differentiation
	GO:0050891	multicellular organismal water homeostasis
	GO:1902622	regulation of neutrophil migration
	GO:0002224	toll-like receptor signaling pathway
	GO:0008630	intrinsic apoptotic signaling pathway in response to DNA damage
	GO:0032649	regulation of interferon-gamma production
	GO:2000106	regulation of leukocyte apoptotic process
12 hours	GO:0050732	negative regulation of peptidyl-tyrosine phosphorylation
	GO:1904892	regulation of receptor signaling pathway via STAT
	GO:0071622	regulation of granulocyte chemotaxis
	GO:0050810	regulation of steroid biosynthetic process
	GO:0045069	regulation of viral genome replication
	GO:0080164	regulation of nitric oxide metabolic process
	GO:0030216	keratinocyte differentiation
	GO:0033561	regulation of water loss via skin
	GO:0062208	positive regulation of pattern recognition receptor signaling pathway
	GO:0062207	regulation of pattern recognition receptor signaling pathway
	GO:0030490	maturation of SSU-rRNA
	GO:1903426	regulation of reactive oxygen species biosynthetic process
	GO:0070098	chemokine-mediated signaling pathway
	GO:0070555	response to interleukin-1
	GO:0098586	cellular response to virus
	GO:0009185	ribonucleoside diphosphate metabolic process
	GO:0071456	cellular response to hypoxia
	GO:1903578	regulation of ATP metabolic process
	GO:0045089	positive regulation of innate immune response
	GO:0006606	protein import into nucleus
	GO:0071675	regulation of mononuclear cell migration
	GO:1901216	positive regulation of neuron death
1 day	GO:0050672	negative regulation of lymphocyte proliferation
	GO:0016052	carbohydrate catabolic process
	GO:0046939	nucleotide phosphorylation
	GO:0072593	reactive oxygen species metabolic process
	GO:0007259	receptor signaling pathway via JAK-STAT
	GO:0071402	cellular response to lipoprotein particle stimulus
	GO:1900371	regulation of purine nucleotide biosynthetic process
	GO:0035335	peptidyl-tyrosine dephosphorylation
	GO:2000404	regulation of T cell migration
	GO:0070229	negative regulation of lymphocyte apoptotic process
	GO:0014066	regulation of phosphatidylinositol 3-kinase signaling
	GO:0010883	regulation of lipid storage
	GO:0043280	positive regulation of cysteine-type endopeptidase activity involved in apoptotic process
	GO:0006911	phagocytosis, engulfment

Peak time	GO ID	Biological Process
	GO:0032088	negative regulation of NF-kappaB transcription factor activity
	GO:0071456	cellular response to hypoxia
	GO:0001776	leukocyte homeostasis
	GO:1903201	regulation of oxidative stress-induced cell death
	GO:0045446	endothelial cell differentiation
	GO:0097191	extrinsic apoptotic signaling pathway
	GO:1903578	regulation of ATP metabolic process
	GO:2000117	negative regulation of cysteine-type endopeptidase activity
	GO:1905954	positive regulation of lipid localization
	GO:0031341	regulation of cell killing
	GO:0015748	organophosphate ester transport
	GO:0071356	cellular response to tumor necrosis factor
3 days	GO:0009620	response to fungus
	GO:0031343	positive regulation of cell killing
	GO:0045621	positive regulation of lymphocyte differentiation
	GO:0002707	negative regulation of lymphocyte mediated immunity
	GO:0043030	regulation of macrophage activation
	GO:0006690	icosanoid metabolic process
	GO:0050891	multicellular organismal water homeostasis
	GO:0009132	nucleoside diphosphate metabolic process
	GO:0006090	pyruvate metabolic process
	GO:0062207	regulation of pattern recognition receptor signaling pathway
	GO:0043331	response to dsRNA
	GO:0045123	cellular extravasation
	GO:0070232	regulation of T cell apoptotic process
	GO:0001523	retinoid metabolic process
	GO:0032642	regulation of chemokine production
	GO:0036294	cellular response to decreased oxygen levels
	GO:0071347	cellular response to interleukin-1
	GO:0072593	reactive oxygen species metabolic process
	GO:1903201	regulation of oxidative stress-induced cell death
	GO:0043123	positive regulation of I-kappaB kinase/NF-kappaB signaling
	GO:0006665	sphingolipid metabolic process
	GO:1905954	positive regulation of lipid localization
	GO:1901136	carbohydrate derivative catabolic process
5 days	GO:0032479	regulation of type I interferon production
	GO:0032660	regulation of interleukin-17 production
	GO:1903900	regulation of viral life cycle
	GO:0048525	negative regulation of viral process
	GO:1905521	regulation of macrophage migration
	GO:0062208	positive regulation of pattern recognition receptor signaling pathway
	GO:0030574	collagen catabolic process
	GO:0032731	positive regulation of interleukin-1 beta production
	GO:0090199	regulation of release of cytochrome c from mitochondria
	GO:0014066	regulation of phosphatidylinositol 3-kinase signaling
	GO:0009593	detection of chemical stimulus
	GO:0032757	positive regulation of interleukin-8 production
	GO:2000379	positive regulation of reactive oxygen species metabolic process

Peak time	GO ID	Biological Process
	GO:0071347	cellular response to interleukin-1
	GO:0031638	zymogen activation
	GO:0006029	proteoglycan metabolic process
	GO:0071456	cellular response to hypoxia
	GO:0045089	positive regulation of innate immune response
	GO:1901216	positive regulation of neuron death
	GO:0003073	regulation of systemic arterial blood pressure
7 days	GO:1902624	positive regulation of neutrophil migration
	GO:0010712	regulation of collagen metabolic process
	GO:0002275	myeloid cell activation involved in immune response
	GO:0014066	regulation of phosphatidylinositol 3-kinase signaling
	GO:0045071	negative regulation of viral genome replication
	GO:1903900	regulation of viral life cycle
	GO:2000107	negative regulation of leukocyte apoptotic process
	GO:0002224	toll-like receptor signaling pathway
	GO:0002688	regulation of leukocyte chemotaxis
	GO:0030510	regulation of BMP signaling pathway
	GO:0032760	positive regulation of tumor necrosis factor production
	GO:0006665	sphingolipid metabolic process
10 days	GO:0030199	collagen fibril organization
	GO:0051785	positive regulation of nuclear division
	GO:0048771	tissue remodeling
	GO:0045766	positive regulation of angiogenesis
	GO:0060541	respiratory system development

4.4 Reduction from original GO lists

The following table shows how each stage of the shortlisting pipeline narrowed down the GO term list length for each dataset. We see that for the top 100 DE genes in mouse and human, for some of the time points, our original GO list is longer than our gene list. This is due to a large amount of redundancy in the GO list resulting from overrepresentation analysis, which overwhelmingly occurs during the beginning of the inflammation stage; we will address this in the next section.

Experiment	Peak time	Length of original GO list	After representation filtering	After similarity filtering
Mouse top 100	6 hours	95	10	6
	12 hours	177	10	6
	1 day	213	18	11
	3 days	124	12	4
	5 days	6	1	1
	7 days	14	1	1
	10 days	25	3	3
Mouse top 1000	6 hours	136	14	9
	12 hours	483	49	22
	1 day	684	69	26
	3 days	518	52	23
	5 days	390	39	20
	7 days	267	27	12
	10 days	71	7	5
Human top 100	3 days	399	39	18
	7 days	117	10	8
	14 days	112	12	8
	21 days	133	14	12

Figure 6: Length of original GO term list obtained from overrepresentation analysis, after representation filtering, and after similarity filtering for each dataset studied.

5 Discussion: wound healing application

In this section, we analyze the resulting shortlists for each dataset in the context of the stages of wound healing by defining each GO term and discussing how its function may fit in to the stage(s) most likely occurring at the given time point as well as within the overall process of wound healing. We speculate how these different processes work together to perform known functions in each stage and why they may be significant.

5.1 Mouse data, top 100 DE genes

6 hours post-wounding, we expect that the wounded tissue is either still undergoing hemostasis or is beginning the inflammation stage. The resulting shortlist of GO terms at this time point towards inflammation. Leukocytes, or white blood cells, are a key part of the inflammatory response. Leukocyte cell-cell adhesion refers to the lining of blood vessels by leukocytes to protect against invading substances at the wound site. Lymphocytes and neutrophils are types of mononuclear cells, which are types of leukocytes, and thus also migrate to the wound site. All these immune cells play important roles in the body's initial defense of the wound site from potential pathogens that might threaten to leak in. Interleukin-8 attracts and activates neutrophils in the inflammatory regions. Reactive oxygen species play a role in inflamma-

tory regulation, as oxidative stress promotes migration of inflammatory cells across the endothelial barriers of wound tissue. Several of the same process continue 12 hours post-wounding, with the addition of interleukin-10 production, which is another type of immune cell with anti-inflammatory properties, and the chemokine-mediated signaling pathway, which stimulates migration of leukocytes to the wound site.

One day post-wounding, we see a continuation of many processes from the first 6 and 12 hours, as the wound is in the midst of the inflammation stage. The acute inflammatory response GO term reflects this overall process. Immune cells continue to migrate to the wound site. Chemotaxis specifically refers to migration along a gradient. Phagocytosis is the process of macrophage activation. Some subtypes of macrophages are leukocytes that eat substances deemed to be dangerous (a potentially dangerous biosubstance is one that does not have on its surface proteins specific to healthy body cells). Toll-like receptors also recognize pathogen associated molecular patterns and defend the body against them, playing an important role in inflammation as well as proliferation. The negative regulation of leukocyte apoptosis is a process that stops leukocytes from committing programmed cell suicide before they are ready to, as they are still needed during this stage. We also see negative regulation of viral genome replication; although we assume there are no viruses at the wound site as the mouse wounds are sterile, defending against viral replication is a part of the immune response and thus may still be activated even if it is not needed. This occurs because genes often perform more than one function or are part of several different pathways, so some highly differentially expressed genes could also be involved in tangentially related inflammatory or immune responses.

On day 3, inflammation continues with the regulation and migration of immune cells and phagocytosis, again all part of the acute inflammatory response. During the last half of the experiment, days 5, 7, and 10, the wound undergoes proliferation and maturation, completing the healing process. Collagen fibril organization occurs, a process by which collagen, the most abundant protein

in the body, rebuilds the wounded skin. On day 10, new blood vessels form via angiogenesis and vasculature development, finalizing the repair of the wounded skin tissue.

We observe that overall, the shortlists obtained from considering the top 100 genes only are quite short, especially for the later days of the experiment. Only one GO term shows up for days 5 and 7, and only 3 terms result for day 10. Compare this to list sizes of 6-12 GO terms for each time point up to day 1, and 4 GO terms for day 3. While the inflammatory/immune response is quite complex and involves many different types of cells, proteins, and regulating factors, proliferation and maturation mostly cover the rebuilding of wounded tissue. It may be the case that less significant processes occur during the proliferation and maturation phases of wound healing, however, there is little else to support this claim. Perhaps a more likely potential reason for this discrepancy in list lengths is that some highly differentially expressed genes may not necessarily correlate with highly significant processes. More research would be needed to investigate this claim as we do not yet have a way of measuring this correlation.

5.2 Mouse data, top 1000 DE genes

Analyzing the top 1000 differentially expressed genes gives us some of the same processes we saw in the top 100, with the addition of many higher-level processes at each time point. However, some of the processes included in the top 100 results do not show up in the top 1000 results, likely due to the mismatch in specificity. Each shortlist is substantially longer as well due to the larger list of genes. In this section, we will discuss the shortlists for each time point, focusing on the contrasting processes, and analyze how they each fit into the big picture of the wound healing process.

6 hours post-wounding, immune cell production, migration, and apoptosis related terms show up as expected. One such term that we did not previously see is regulation of interferon-gamma production; interferon-gamma cells are crucial for the immune response and differentiate into T cells, another impor-

tant type of immune cell. Another new term is keratinocyte differentiation; keratinocytes are responsible for restoring the epidermis after injury and their differentiation is critical for epidermal stratification, or the formation of a barrier in the skin to protect the body, which may be a part of hemostasis. We also see a term related to water homeostasis, or the regulation of the amount of water present in the body. This may relate to preventing additional fluid loss due to bleeding and thus be tangentially related to hemostasis, which we expect to see during this time point.

The shortlist for 12 hours post-wounding is significantly longer as well and introduces many new processes as well as keeping some familiar ones, such as response to interleukin proteins and immune cell migration. A couple of terms for the regulation of pattern recognition receptor signaling show up; pattern recognition has to do with recognizing molecules typically found in pathogens and is thus a part of the immune response. Tyrosine phosphorylation has many functions such as cell adhesion, proliferation, migration, differentiation, gene regulation, and angiogenesis; this process is downregulated, potentially indicating a change in state of the tissue in which some of these functions are no longer needed, however the vagueness of this term makes it difficult to say which ones. Signal transducer and activator of transcription (STAT) mediates immune cells as well. Neuron death is upregulated, likely due to the need to kill nerves that were damaged during wounding and prepare for innervation, the formation of new nerves. The cellular response to hypoxia is likely a result of the reactive oxygen species synthesis which causes oxidative stress on cells. Hypoxia is a metabolic shift to a more active state that allows cells to produce necessary components for proliferation. Finally, we see a few tangentially related processes such as regulation of viral genome replication and steroid synthesis. As we saw with the top 100 list, viral regulation terms show up due to the fact that they may be activated during the immune response even if there is no virus present. Steroids are cell membrane components and signaling molecules, so they may have some higher level function related to immune signaling pathways.

On day 1 post-wounding, we know that the inflammation stage is still ongoing but several processes begin to slow down in preparation for the advanced stages. Lymphocyte proliferation is downregulated but so is its apoptosis; these immune cells are still needed even if not in large amounts. Leukocyte homeostasis refers to regulating the proliferation and elimination of leukocytes following the immune response, possibly limiting the number of leukocytes present in the wound site as well. Interestingly, we see that cysteine-type endopeptidase activity is both upregulated and downregulated. Cysteine-type endopeptidase cuts the internal peptide bonds in or around cysteine, which is an amino acid crucial for the production of collagen; its positive and negative regulation may indicate the breakage of damaged collagen as well as the beginning of collagen rebuilding, however, it is likely too early in the wound healing process for rebuilding as that is part of the proliferation stage. Phosphatidylinositol 3-kinase signaling regulates cell proliferation and apoptosis as well as cytoskeletal rearrangement, the latter of which is also necessary for breaking down the damaged tissue to prepare for rebuilding. Lipoproteins also play a role in the inflammatory process, interacting with immune cells to modulate the immune response, hence the cellular response to lipoprotein particle stimulus term. The receptor signaling pathway via JAK-STAT conveys a signal to trigger a change in the activity or state of a cell, which could also have many implications in terms of the regulation of immune cells. Endothelial cells constitute the innermost layer of blood and lymphatic vessels, and their differentiation is necessary for the formation of new vessels from pre-existing broken ones to prepare for the process of angiogenesis later on.

We expect that day 3 post-wounding covers the later segments of the inflammatory response, and see that at this point the wound tissue is still utilizing lymphocytes, macrophages, T cells and chemokines, however some of these responses seem to be slowing down as evidenced by lymphocyte immune response downregulation and T cell apoptosis. The response to interleukin and hypoxia is still occurring as well. NF-kappaB also plays a regulatory role in the cellular response to pro-inflammatory cytokines. We also see many new terms including retinoid, sphingolipid and eicosanoid metabolism, extravasation, and

carbohydrate catabolism. Retinoids help regulate epithelial cell growth, part of proliferation. Eicosanoids have many functions including the inflammatory response and perception of pain. Sphingolipids play roles in signal transduction and cell recognition and are part of the immune process. Extravasation is the emigration of cells from the blood stream through the vascular endothelium into the tissue and occurs in leukocytes, which are still being recruited to the wound site. Carbohydrates are essential for skin cell energy which is needed more excessively during wound healing, and also help to regulate cell adhesion, migration, and proliferation.

By day 5, the mouse tissue is undergoing proliferation, and in the top 100 genes, the only significant term that showed up was for collagen organization. However, the results for the top 1000 show many GO terms that may relate to both inflammation and proliferation, but do not show collagen organization. Inflammation-related terms include regulation of interleukins, interferon, macrophages, pattern recognition, and innate immune response, response to reactive oxygen species, and collagen catabolism, all of which have showed up during earlier stages. Broader terms include zymogen activation, the process of creating an active enzyme from the inactive biomolecule zymogen, and release of cytochrome c, part of the apoptotic process. Terms that could potentially be related to proliferation include regulation of arterial blood pressure, since if blood pressure is too high, cells don't get enough oxygen and thus cannot regenerate, and proteoglycan metabolism, since proteoglycans are a major component of the extracellular matrix, which plays a key role in the proliferation stage of wound healing. Phosphatidylinositol 3-kinase signaling is again present, we recall from day 1 that this term has many functions, including the regulation of proliferation, apoptosis, and cytoskeletal rearrangement. Clearly the proliferation regulation, and the cytoskeletal rearrangement as well, should be part of the proliferation stage.

We expect proliferation-related terms to show up during day 7 as well, with the top 100 results for this day being the same as those for day 5. Some inflammatory processes are still occurring, including neutrophil migration, toll-like

receptor signaling, leukocyte chemotaxis, sphingolipid metabolism, and tumor necrosis factor production. Myeloid cell activation involved in immune response shows up as well; myeloid is bone marrow tissue which is an important part of the immune system. Proliferation related processes include collagen metabolism, which is needed to produce or replace collagen for proliferation and wound tissue rebuilding, and regulation of BMP (bone morphogenetic protein) signaling pathway, which play a role in the regulation of cell proliferation. We also see phosphatidylinositol 3-kinase signaling again as with day 5.

Day 10 post-wounding, the wound tissue is undergoing the maturation stage. As expected from our top 100 results, we see collagen fibril organization and angiogenesis, important processes that occur in order to rebuild the wound tissue, as well as tissue remodeling. Positive regulation of nuclear division also occurs; it is likely that this term relates to the cell proliferation that must occur in order to rebuild the wound site.

We also see several processes in the shortlists for some of the time points that do not seem to have anything to do with wound healing, which have been omitted from this discussion. There are several reasons why these processes might have survived the filtering process despite their irrelevance. For one, some genes have many different functions, many of which are unknown, and the annotation process may not cover all such functions. The Gene Ontology also does not have the capability to produce results according to context, so it may output all terms annotated to a given gene regardless of relevance in the current context, or which pathways are expressed. Perhaps also some genes are being expressed during the time of the experiment that are not directly related to wound healing, but happen tangentially due to links we are thus far unaware of. Finally, it is always possible that genes have been incorrectly annotated in the database. While we have no control over the latter issue, further edits of GO should help to fix potentially incorrect annotations.

5.3 Human data, top 100 DE genes

Our first data point for human wound healing is day 3 post-wounding, at which point the wound tissue is undergoing inflammation. Recalling that human wounds take longer to heal than mouse wounds due to their larger size, and the fact that the mouse wounds were sterile while human wounds were not, we expect the process to be slower and for the inflammatory/immune response to go on for quite some time. As expected, nearly all resulting terms on day 3 were inflammation related. These results include some familiar terms we saw in the mouse data such as T cell activation, neutrophil chemotaxis, chemokine signaling, interleukin production, response to cytokines, and leukocyte apoptosis. The less specific terms acute inflammatory response and regulation of type 2 immune response, which maintains metabolic homeostasis and regulates tissue repair after injury, occur as well. Other inflammation related terms that come up are as follows. Leukocyte degranulation, the release of immune cell contents; response to heat, as heat generation is a part of the immune response; antigen processing and presentation of exogenous peptide antigen, meaning that an antigen presenting cell expresses a peptide antigen to defend against pathogens coming from outside the organism; and nitric oxide metabolism, which induces inflammation and vasodilation involved in the immune response by cytokine activated macrophages. Finally, disassembly of the extracellular matrix and cell junctions occur. Cell junctions are connections between two cells or between a cell and the extracellular matrix. While these functions are not directly related to inflammation, they likely occur in preparation for later stages in which the extracellular matrix components must proliferate and rebuild over the wound site.

Day 7 post-wounding, inflammation is still in full swing with the migration of leukocytes and neutrophils to the wound site. The acute phase response, or a nonspecific reaction to injury or inflammation, also occurs; the most important sources of acute phase proteins are macrophages and monocytes. Extracellular matrix components continue to be broken down as signified by the collagen catabolic process. Some proliferation associated processes are also expressed at this time, namely, upregulation of vascular endothelial growth factor, a

signaling protein that promotes growth of new blood vessels, and keratinization, the process by which epithelial cells produce large amounts of keratin filaments, a tough structural protein used in skin repair.

The processes expressed on day 14 can likely be characterized as still in the early proliferation phase. Homotypic cell-cell adhesion may refer to the formation of an "immunological synapse" between immune cells, an inflammatory response, or forming new cellular structures, a proliferative activity. Collagen catabolism is still occurring at this time as well. The downregulation of cellular response to growth factor may refer to slowing down cell proliferation or immune cell production as inflammation ends, however we do not have enough information as to which types of cells are undergoing this response. Upregulation of epithelial to mesenchymal transition occurs; mesenchymal cells are cells with the ability to differentiate into any type of smooth muscle, vascular endothelium, connective tissue, blood vessels, or lymphatic tissue. They are also resistant to apoptosis and altered synthesis of extracellular matrix components. The upregulation of smooth muscle cell migration occurs during vascular development in response to injury, and is the recruitment of cells to areas where the vessel wall is being remodeled. These processes are likely related to proliferation as we need new structures to be rebuilt out of the old damaged skin and tissues.

On day 21, we expect that the wound tissue is undergoing proliferation and/or remodeling, depending on the severity of the wound, both of these stages may be long-lasting. However, we still see several inflammation related processes showing up, likely due to the fact that many of these inflammatory responses are very highly upregulated compared to later stage processes, and thus associated genes may take a while to stop appearing in the most highly upregulated lists. Inflammation related processes include leukocyte chemotaxis, interferon gamma production, and downregulation of viral replication. We discussed in our analysis of the mouse data that the latter term is part of the immune response whether or not viruses are actually present in the wound tissue, although since the human wounds are not sterile, the viral response may have

occurred due to actual stimuli. Extracellular matrix disassembly and down-regulation of cytokine-mediated signaling pathway represent processes likely taking place in between inflammation and proliferation. Proliferation related processes include collagen metabolic process, regulation of BMP (bone morphogenetic protein) signaling pathway, which plays a role in cell proliferation, and morphogenesis of branching structures, or the generation of branching blood vessels, lymphatics, nerves, or epithelial tubes, all of which must be rebuilt during the proliferative phase. Finally, tissue remodeling, the main process occurring during maturation, is also expressed. However, we do not expect that the maturation and remodeling phase is finished by 3 weeks post wounding, as this may go on for up to a year in humans.

5.4 Comparing mouse and human top 100 shortlists

In comparing the times and length of onset of mouse and human wound healing stages, we notice that mouse wounds heal much faster than human wounds. While days 0 through 10 in mouse covered each wound healing stage, we need to look at data at least through day 21 in humans to begin reaching the final stage, and even then, healing is not complete for much longer. The reasons for the large differential in wound healing time are that human wounds are larger and take longer to heal than small mouse wounds. Additionally, experiments on mice are done in a sterile lab environment and thus are much less likely to become infected, whereas human wounds are not guaranteed to be sterile.

The shortlists for human are also on average longer than those for mouse. Since the human genome is larger than the mouse genome, there are far more GO terms annotated to human genes, and since a single gene can be annotated to any number of GO terms, we see more significant GO terms when performing overrepresentation analysis. This makes comparison between the two datasets somewhat more complicated. Late-stage human data shows tissue remodeling, an important maturation-related process, which does not come up in the mouse data. We also see more detailed proliferation related processes such as keratinization, smooth muscle cell migration, and development of undifferentiated cells that can produce lymphatics, endothelium, and connective

tissue as well as blood vessels, in human but not in mouse.

However, we still saw many of the same main biological processes come up in both datasets. Inflammatory response is mostly consistent, with migration and regulation of different types of immune cells occurring at the earlier stages in both mouse and human and slowing down during later stages. During the proliferation stage, we see terms related to vasculature development and collagen metabolism/extracellular matrix construction, in both, also as expected. This verifies that many of the high level processes are the same even if the details vary slightly.

5.5 Comparing top 100 and top 1000 in mouse

We noted in the previous section when we introduced testing the method with a larger list size that due to the output of overrepresentation analysis, we expect that significant GO terms for larger list sizes are on average less specific than for smaller lists, especially when selecting for GO terms with higher representation in the dataset. We shifted the biases in computing the representation score to promote higher representation of more specific GO terms, however, our method is not perfect and the resulting GO terms for the top 1000 differentially expressed genes were still largely different from the results for the top 100. We also obtained more GO terms that were seemingly unrelated to wound healing than for the top 100. When we increase the list by an order of magnitude, we include differentially expressed genes that may also code for several functions or be part of several pathways that are unrelated, or only tangentially related, to wound healing. It also may be the case that these GO terms represent functions that do play a role in wound healing that is as of yet undiscovered. Alternatively, including genes with low fold change could add unrelated noise in our results, causing unrelated GO terms to show up. While genes with low fold change may correspond to wound healing related processes that are simply less expressed, they may also represent functions completely unrelated that show up in the microarray in addition as a result of enforcing a lower threshold. Further research will be needed to assess this.

6 Conclusions & further research

In this work, we presented a method for producing shortlists of GO terms from the often superfluous lists of GO terms resulting from standard overrepresentation analysis. This study addressed two outstanding issues with the current state of GO analysis, namely, the inclusion of GO terms that are overly general or too specific to be significantly represented, and the addition of redundant GO terms. We provided a heuristic solution for both of these issues in the form of a pipeline that shortens a GO list using a stage of filtering for each issue; specificity/representation and similarity of GO terms. Applying this method to timeseries transcriptomal profiling of mouse and human skin wounds, we were able to identify some of the significant dynamic biological processes occurring during each stage of wound healing. We discuss the main takeaways from this study and areas that need improvement in this section.

While this work presents progress in computational methods for gene ontology analysis, there remains a need for further analysis and development in future work. One significant such issue is that we do not set an exact threshold for differential expression, but rather rank genes in fold change from highest to lowest and take the top x number of genes. This may result in a different threshold being implicitly set for each day of the experiment. Additionally, amount of upregulation does not necessarily equate to level of significance. For example, many inflammation related genes start out very highly upregulated in the beginning days, and progressively become somewhat less upregulated later on, while genes related to proliferation and maturation may be less upregulated according to measured fold change even if their expression appears to be significant enough from a visual standpoint. This could result in many inflammation related terms showing up in almost all of the time points, while fewer proliferation and maturation related terms show up towards the end than we might expect.

We also see a lack of GO terms related to the hemostasis stage of wound healing. While hemostasis typically only lasts for a few hours post-wounding, we would expect hemostasis-related biological processes to show up at the 6

hour time point in mouse, however, we only see inflammation-related processes at this time. There are several reasons that this may have occurred. First, there are, in general, less GO terms related to hemostasis processes than inflammation, which is a highly involved stage with many subprocesses. Also, processes related to hemostasis may simply be less highly upregulated than those related to inflammation, regardless of their importance.

Fine-tuning the thresholds to reflect how much upregulation is actually significant for each type of gene or stage in wound healing is one potential way to mitigate these issues. However, there are far too many unknowns to reliably test this as we have not yet measured the correlation between amount of upregulation and functional significance. Another, more realistic way we might go about resolving this issue is the integration of our computational gene ontology-based method with other predictive models, namely, an image-based prediction algorithm that maps actual images of wounds to their corresponding stage based on machine learning data. This may help us obtain more accurate data regarding mapping wound healing stages to dynamic biological processes with respect to both gene expression data and real-time wound data.

Another significant limiting factor in the efficacy of this method is the accuracy of the PANTHER database and the GO itself. Although overrepresentation analysis is based on statistical data from many sources, and the database is updated frequently, there are still many unknowns that modern science has yet to provide an answer for. Again, combining this method with other models could help mitigate issues related to the gene analysis side of the wound healing problem.

In future work, the shortlisting pipeline could be improved by automating the bias selection process. As we mentioned in the methods section, we currently choose biases in order to shape our optimization function for representation score as a concave-down quadratic with a maximum in the 10-12 bit range. This resulted in a lot of time being spent in testing which biases produced such a result for each dataset, as the optimal point was slightly different for

each one. This was most noticeable in testing larger gene lists in which the GO terms were more general on average; the biases needed to be changed drastically in order to push the GO terms in the 10-12 bit range to the maximum of the representation score function, else the resulting terms would have much too low information content to be relevant. While exhaustive testing of which biases produced the most accurate results worked well for this study, it may not scale to larger scope experiments. Automating the bias selection may take the form of a program that computes the representation score with a wide variety of biases, plots the results, and applies a best-fit quadratic to the data and computes the maximum. It would then find the b_1 and b_2 that gave a maximum closest to the desired value and whose quadratic fit was most precise.

We could also improve the method with further understanding of the similarity filtering aspect. Currently, we weight the semantic similarity score more than the functional score. As mentioned earlier, this is mainly motivated by the lack of functional score availability for many of the pairs of GO terms in the database, but also due to the fact that the semantic similarity gives us more consistent results; pairs of terms that are very closely related in the GO DAG are very likely to be redundant, while relatively less is known about the functional similarity of annotation frequency. More data on the latter would help us to develop a more balanced understanding of the similarity process as a whole.

From a purely technical perspective, the current method still leaves a lot to be desired. The codes still require some grunt work from the user in terms of entering data into the PANTHER and NaviGO databases, keeping track of the different spreadsheets and data produced by each step of the pipeline, and manually adjusting the dataset to be used as well as customizing biases for each dataset. Integrating the data-entry and online database interface into the code would make the tool much more powerful for collaborative use. Further work on the computer science side of developing the pipeline could greatly improve the value of this method as a whole.

7 References

[1] Alterovitz, Gil, et al. “GO PaD: The Gene Ontology Partition Database.” Nucleic Acids Research, no. suppl_1, Oxford University Press (OUP), Nov. 2006, pp. D322–27. Crossref, doi:10.1093/nar/gkl799.

[2] Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

[3] “Gene Ontology Resource.” Gene Ontology Resource, http://geneontology.org/. Accessed 28 Nov. 2021.

[4] “GEO Accession Viewer.” National Center for Biotechnology Information, https://ncbi.nlm.nih.gov/geo/query/acc.cgi. Accessed 28 Nov. 2021.

[5] “Information Theory - Wikipedia.” Wikipedia, the Free Encyclopedia, Wikimedia Foundation, Inc., 20 May 2001, https://en.wikipedia.org/wiki/Information_theory.

[6] Khatri, P., and S. Draghici. “Ontological Analysis of Gene Expression Data: Current Tools, Limitations, and Open Problems.” Bioinformatics, no. 18, Oxford University Press (OUP), June 2005, pp. 3587–95. Crossref, doi:10.1093/bioinformatics/bti565.

[7] Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. Jan 2019;47(D1):D419-D426.

[8] “The Four Stages of Wound Healing — WoundSource.” WoundSource, 28 Apr. 2016, https://www.woundsource.com/blog/four-stages-wound-healing.

[9] The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. Jan 2021;49(D1):D325-D334.

[10] Seyhan, Attila A. “Lost in Translation: The Valley of Death across Preclinical and Clinical Divide – Identification of Problems and Overcoming Obstacles.” Translational Medicine Communications, no. 1, Springer Science and Business Media LLC, Nov. 2019. Crossref, doi:10.1186/s41231-019-0050-7.

[11] “WebGestalt (WEB-Based GEne SeT AnaLysis Toolkit).” WebGestalt (WEB-Based GEne SeT AnaLysis Toolkit), http://www.webgestalt.org/. Accessed 6 Dec. 2021.

[12] Wei, Qing, et al. “NaviGO: Interactive Tool for Visualization and Functional Similarity and Coherence Analysis with Gene Ontology.” BMC Bioinformatics, no. 1, Springer Science and Business Media LLC, Mar. 2017. Crossref, doi:10.1186/s12859-017-1600-5.

[13] Yon Rhee, Seung, et al. “Use and Misuse of the Gene Ontology Annotations.” Nature Reviews Genetics, no. 7, Springer Science and Business Media LLC, May 2008, pp. 509–15. Crossref, doi:10.1038/nrg2363.

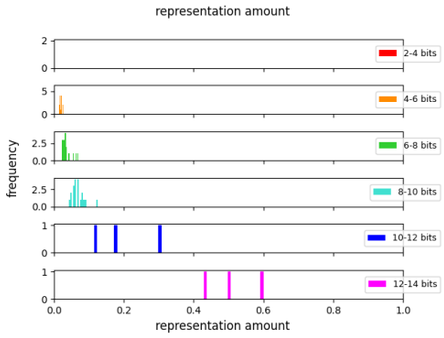
[14] Zlobina, Ksenia, et al. Transcriptomic Time Series Analysis in Wound Healing: Challenges and Perspectives on Data Interpretation. Research Square Platform LLC, Oct. 2021. Crossref, doi:10.21203/rs.3.rs-929173/v1.

8 Appendix

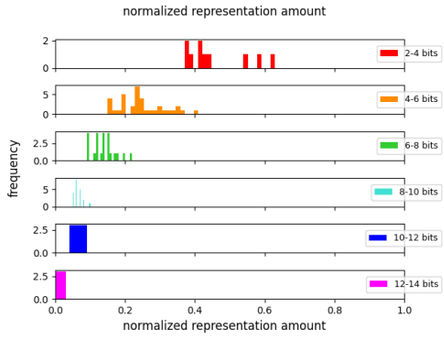
All figures generated in this work – histograms and scatter plots from the representation filtering stage and similarity network graphs from the similarity filtering stage of the pipeline – are presented in this appendix. The histograms, scatter plots, and similarity graphs were produced for the top 100 and top 1000 differentially expressed (DE) mouse genes at each time point, and the top 100 DE human genes at each time point. The histograms, labeled (a) and (b) in each figure, illustrate the distributions of the representation amount and normalized representation amount of GO terms, respectively, for successive intervals of information content in bits. The scatter plots, labeled (c) in each figure, show representation amount, normalized representation amount, and representation score versus bits, where each dot represents a single GO term. The similarity graphs, labeled (d) in each figure, show which GO terms were determined to be sufficiently similar to one another. Figures with no similarity graph were datasets for which no pairs of GO terms met the similarity criteria used in this study. Refer to section 3.2 for full descriptions on how each of these figures are produced.

8.1 Mouse data, top 100 DE genes

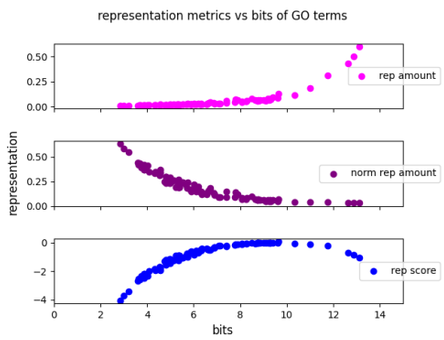
Gene expression data was taken at 6 and 12 hours, and days 1, 3, 5, 7, and 10 post-wounding of mouse skin. Figures corresponding to the top 100 DE genes at each time point are shown in this section.



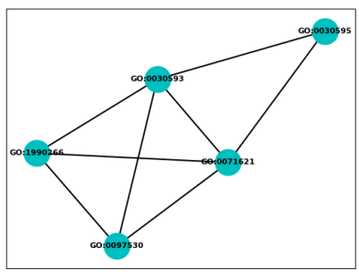
(a) Representation amount



(b) Normalized representation amount

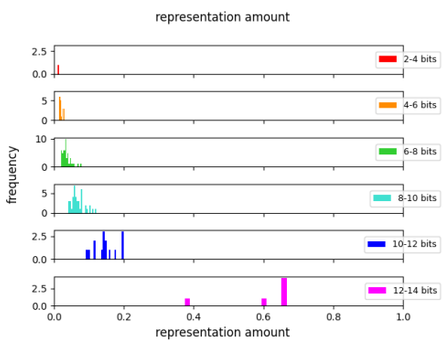


(c) Representation metrics vs bits

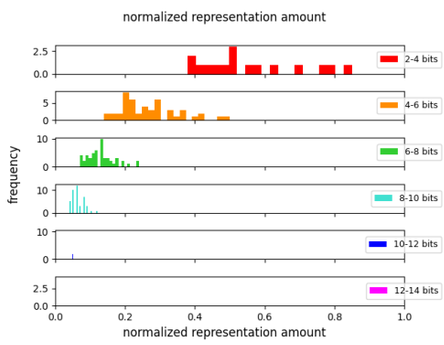


(d) Similarity graph

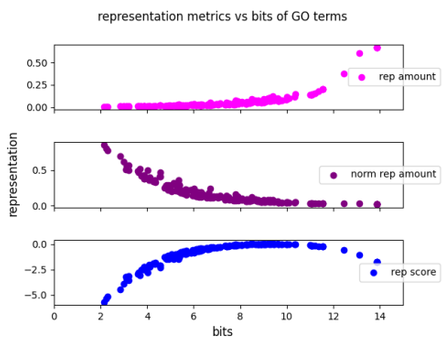
Figure 7: Mouse top 100, 6 hours



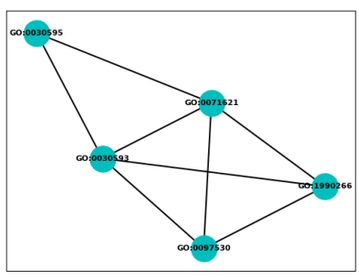
(a) Representation amount



(b) Normalized representation amount

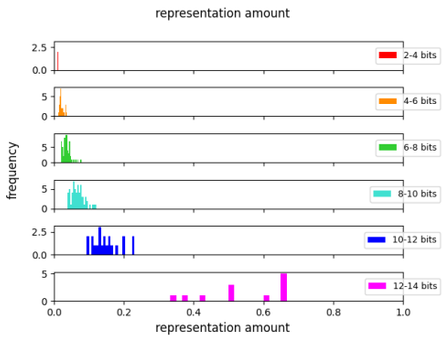


(c) Representation metrics vs bits

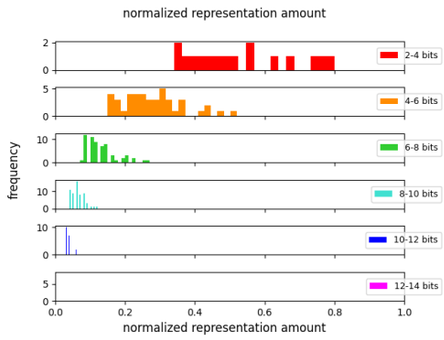


(d) Similarity graph

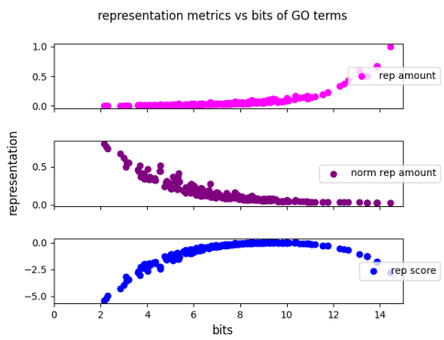
Figure 8: Mouse top 100, 12 hours



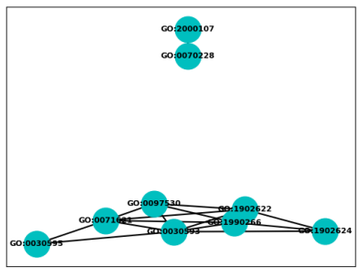
(a) Representation amount



(b) Normalized representation amount

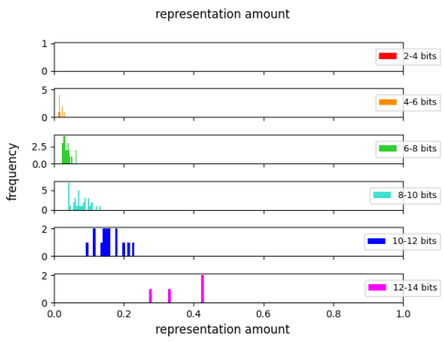


(c) Representation metrics vs bits

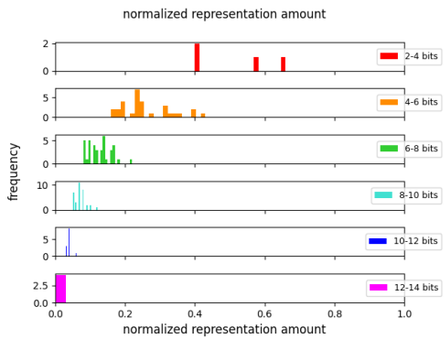


(d) Similarity graph

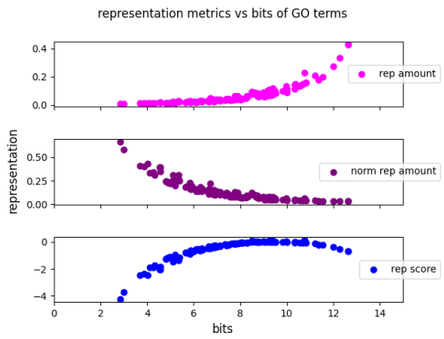
Figure 9: Mouse top 100, 1 day



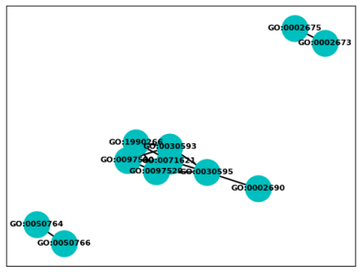
(a) Representation amount



(b) Normalized representation amount

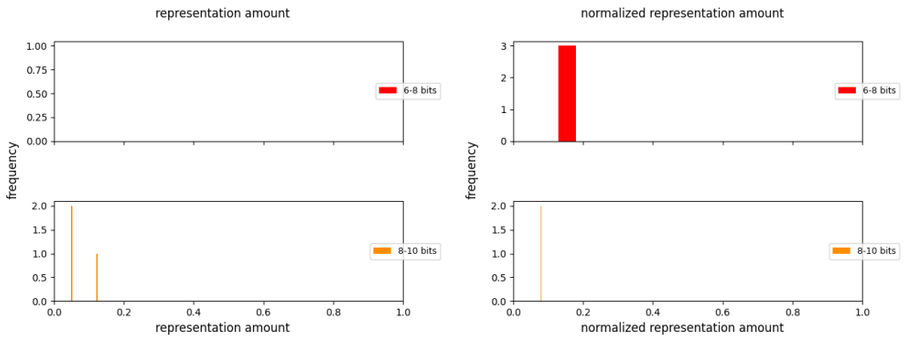


(c) Representation metrics vs bits

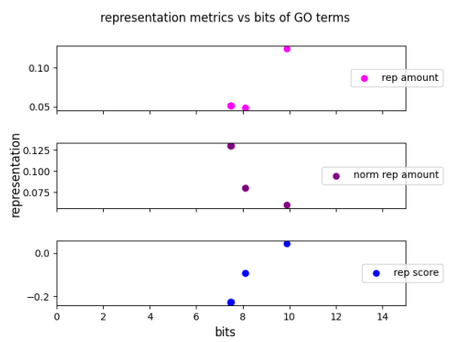


(d) Similarity graph

Figure 10: Mouse top 100, 3 days

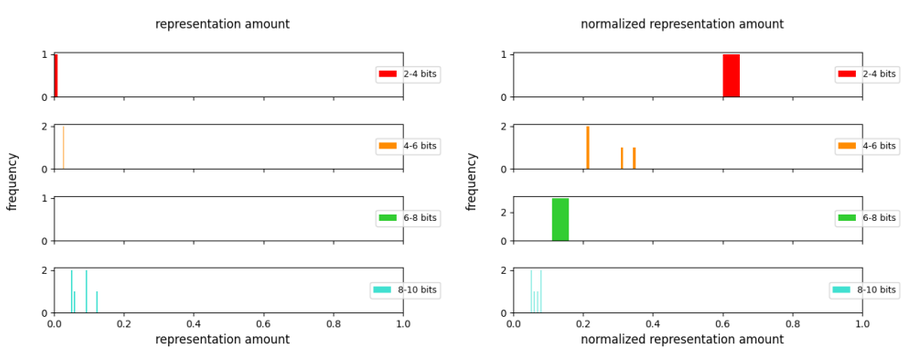


(a) Representation amount (b) Normalized representation amount

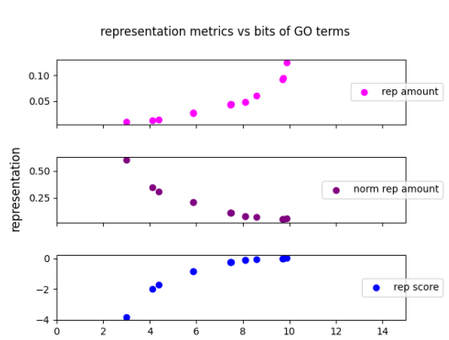


(c) Representation metrics vs bits

Figure 11: Mouse top 100, 5 days

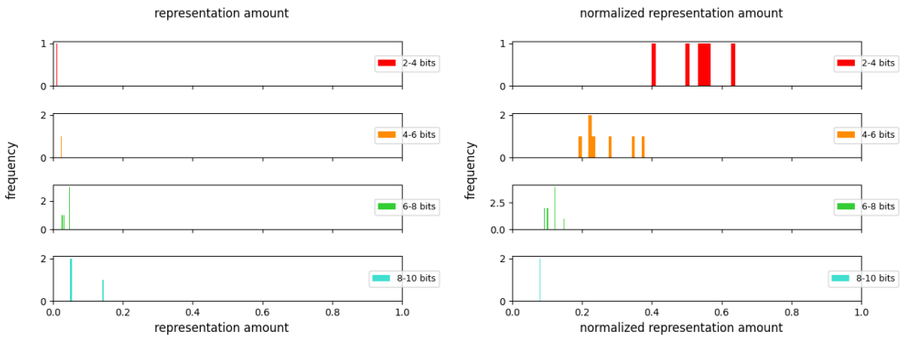


(a) Representation amount (b) Normalized representation amount



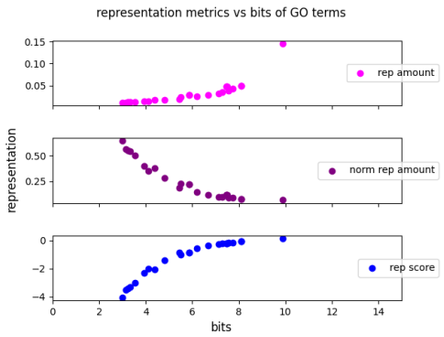
(c) Representation metrics vs bits

Figure 12: Mouse top 100, 7 days



(a) Representation amount

(b) Normalized representation amount

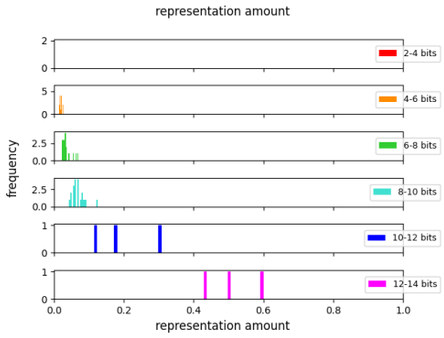


(c) Representation metrics vs bits

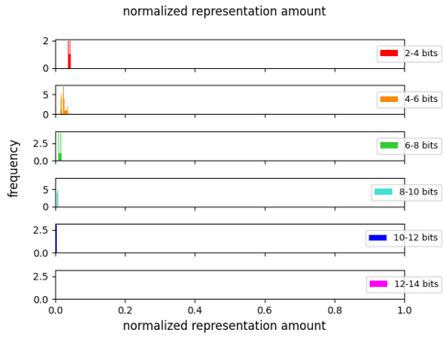
Figure 13: Mouse top 100, 10 days

8.2 Mouse data, top 1000 DE genes

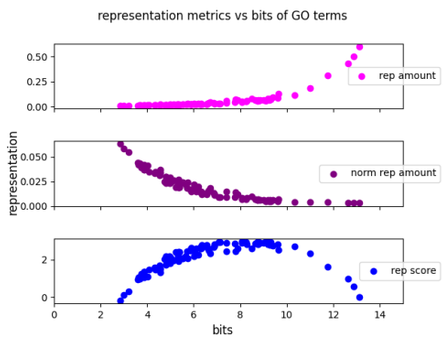
Gene expression data was taken at 6 and 12 hours, and days 1, 3, 5, 7, and 10 post-wounding of mouse skin. Figures corresponding to the top 1000 DE genes at each time point are shown in this section.



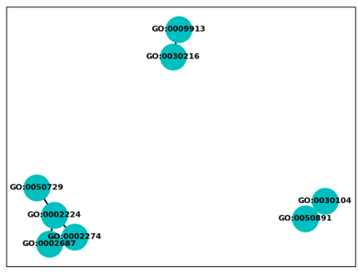
(a) Representation amount



(b) Normalized representation amount

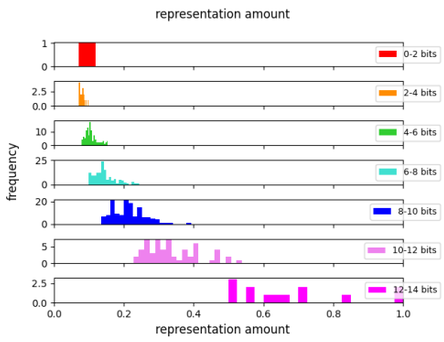


(c) Representation metrics vs bits

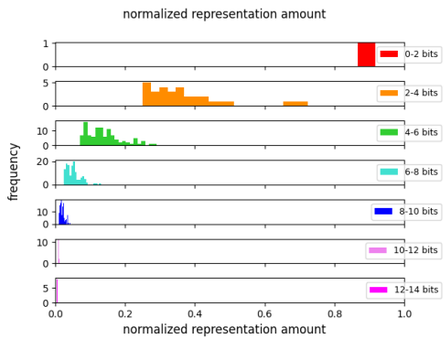


(d) Similarity graph

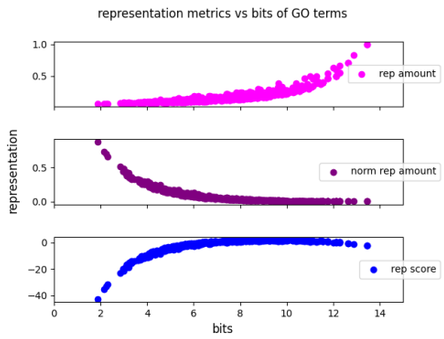
Figure 14: Mouse top 1000, 6 hours



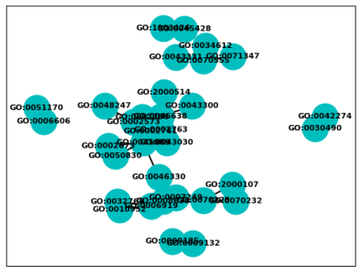
(a) Representation amount



(b) Normalized representation amount

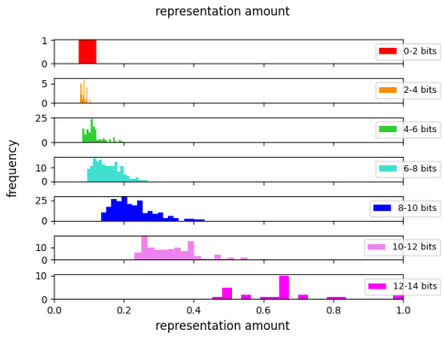


(c) Representation metrics vs bits

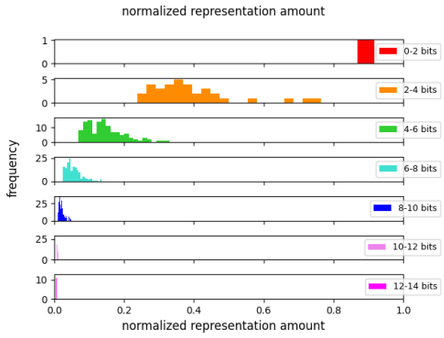


(d) Similarity graph

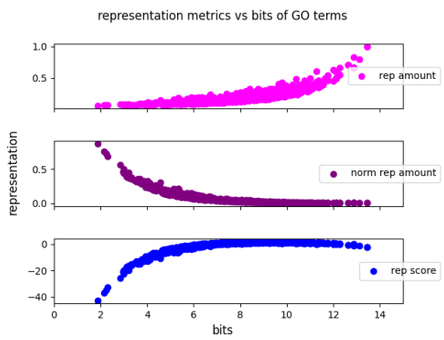
Figure 15: Mouse top 1000, 12 hours



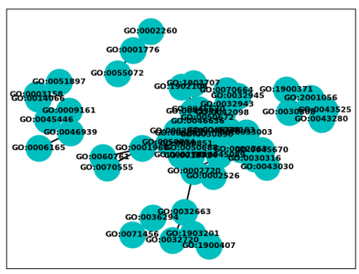
(a) Representation amount



(b) Normalized representation amount

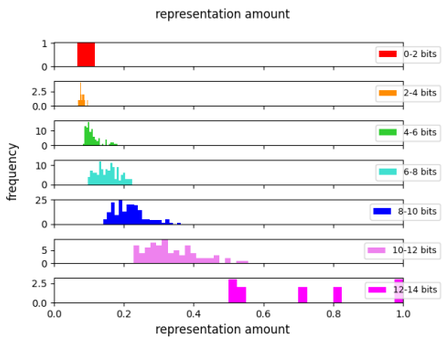


(c) Representation metrics vs bits

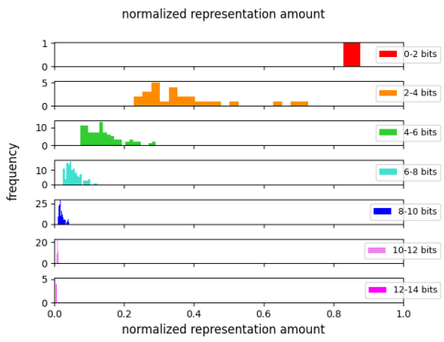


(d) Similarity graph

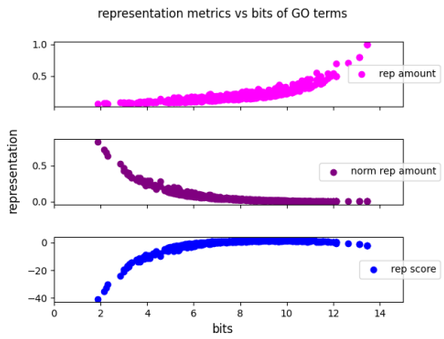
Figure 16: Mouse top 1000, 1 day



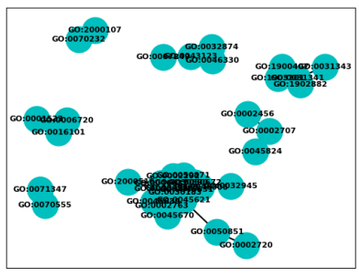
(a) Representation amount



(b) Normalized representation amount

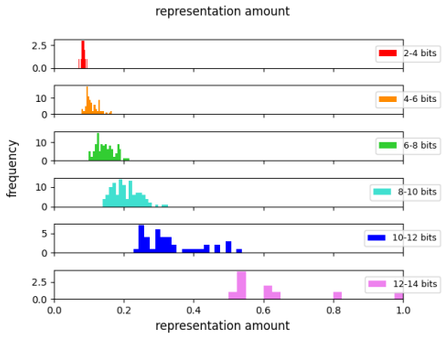


(c) Representation metrics vs bits

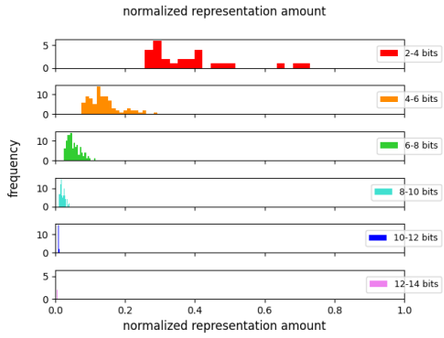


(d) Similarity graph

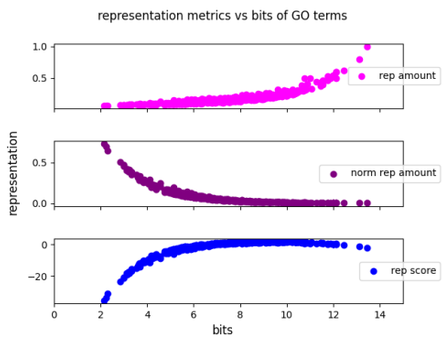
Figure 17: Mouse top 1000, 3 days



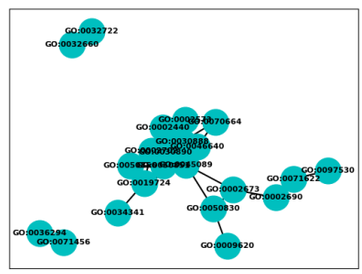
(a) Representation amount



(b) Normalized representation amount

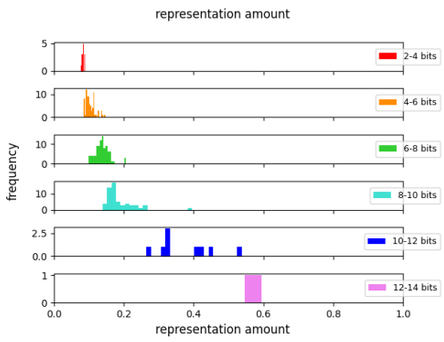


(c) Representation metrics vs bits

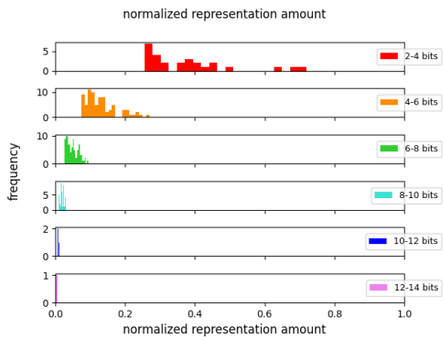


(d) Similarity graph

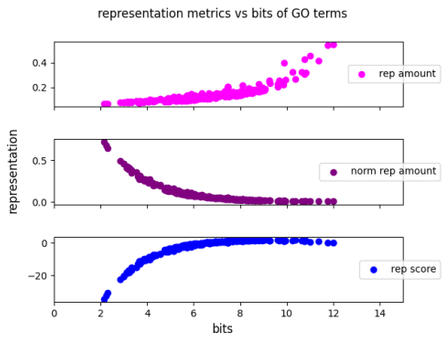
Figure 18: Mouse top 1000, 5 days



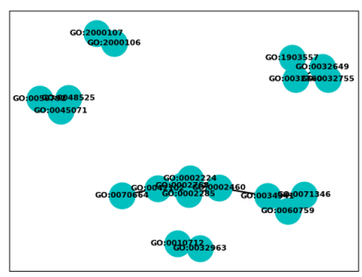
(a) Representation amount



(b) Normalized representation amount

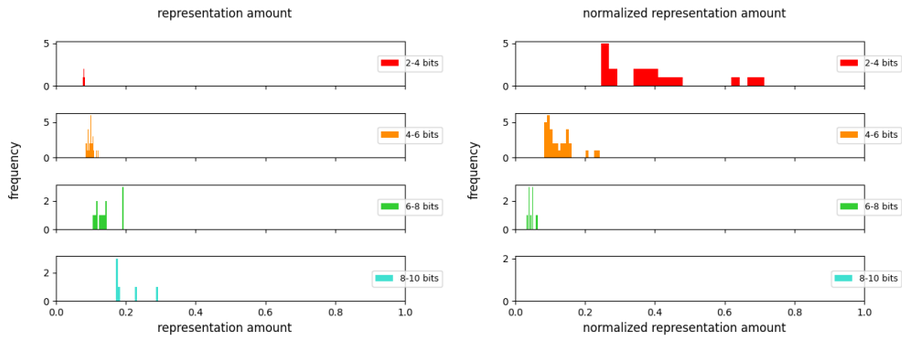


(c) Representation metrics vs bits



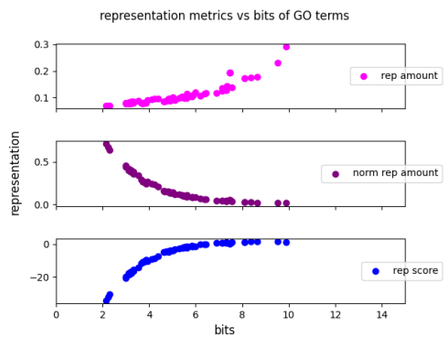
(d) Similarity graph

Figure 19: Mouse top 1000, 7 days

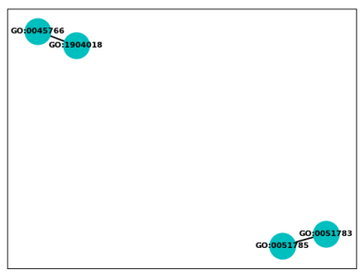


(a) Representation amount

(b) Normalized representation amount



(c) Representation metrics vs bits

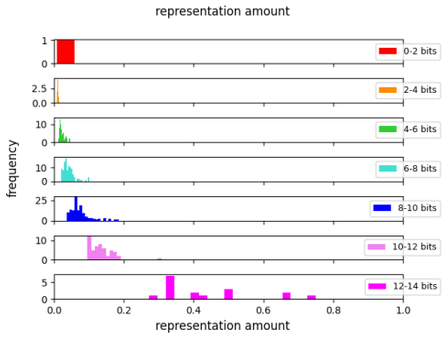


(d) Similarity graph

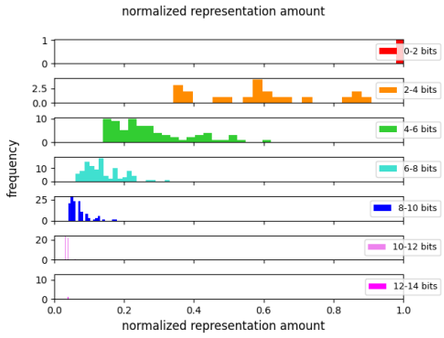
Figure 20: Mouse top 1000, 10 days

8.3 Human data, top 100 DE genes

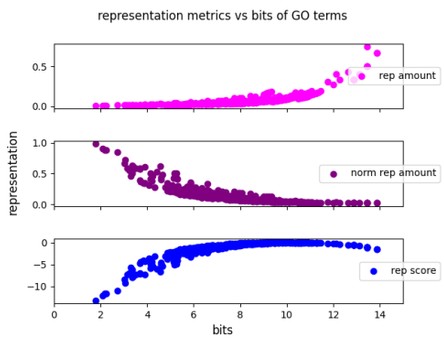
Gene expression data was taken at days 3, 7, 14, and 21 post-wounding of human skin. Figures corresponding to the top 100 DE genes at each time point are shown in this section.



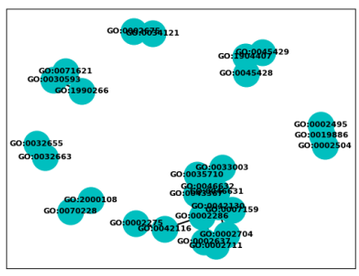
(a) Representation amount



(b) Normalized representation amount

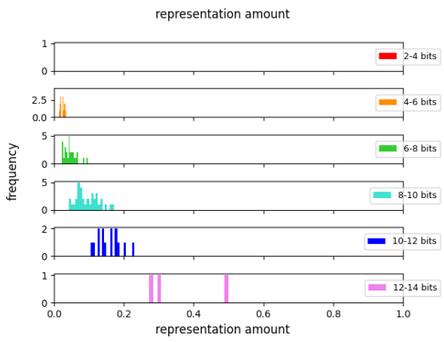


(c) Representation metrics vs bits

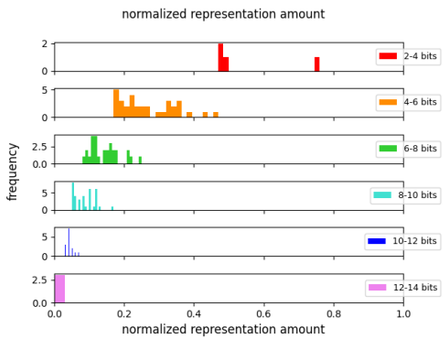


(d) Similarity graph

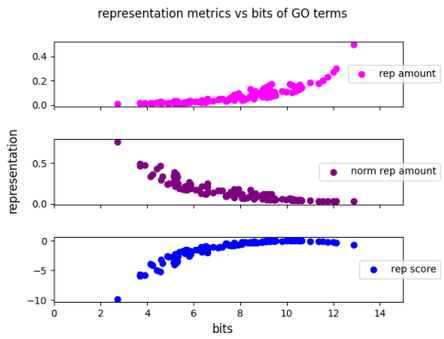
Figure 21: Human top 100, 3 days



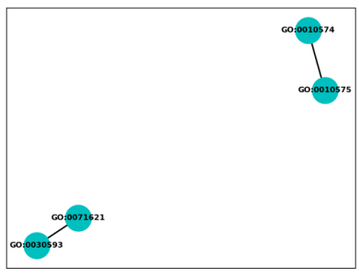
(a) Representation amount



(b) Normalized representation amount

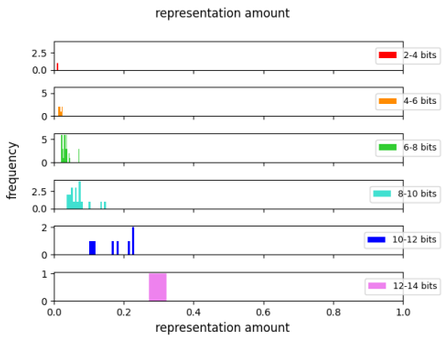


(c) Representation metrics vs bits

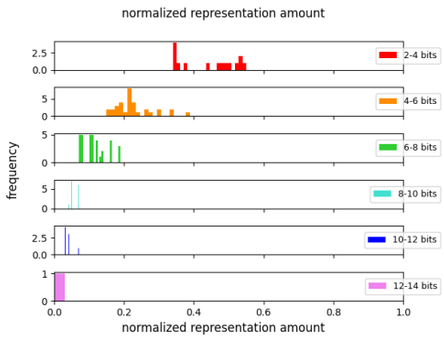


(d) Similarity graph

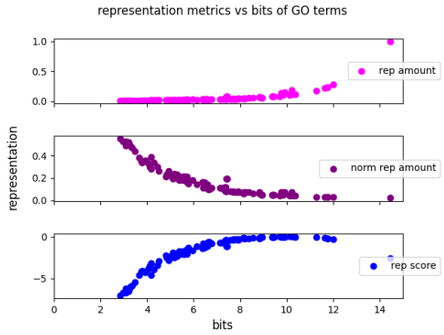
Figure 22: Human top 100, 7 days



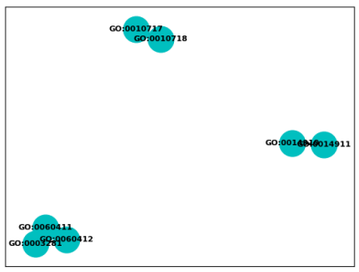
(a) Representation amount



(b) Normalized representation amount

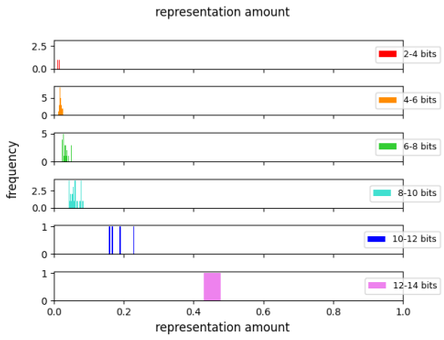


(c) Representation metrics vs bits

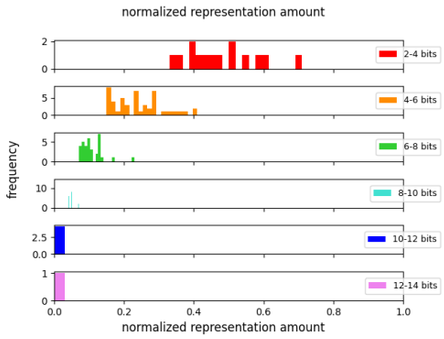


(d) Similarity graph

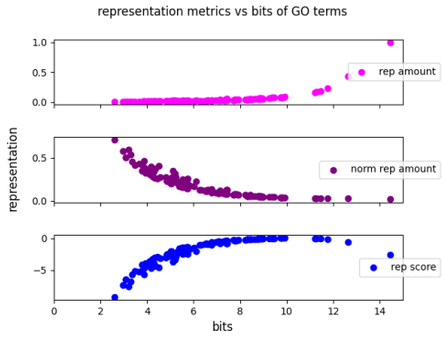
Figure 23: Human top 100, 14 days



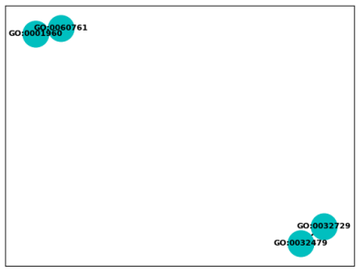
(a) Representation amount



(b) Normalized representation amount



(c) Representation metrics vs bits



(d) Similarity graph

Figure 24: Human top 100, 21 days