Technical Note General Word Limit: 3000 Words

Target Journal, Impact Factor

PeerJ, 2.7

Bioinformatics, 5.8

BMC Bioinformatics, 2.9

PLOS Computational Biology, 3.8

Microbial Genomics, 3.9

BMC Genomics, 3.5

Scientific Reports, 3.8

Technical Notes

Need to add genome assessments, number of genes found with prodigal, remove outliers…

Add references…

Add new heatmap-like figure…

**Title (12)**

GOOP: A Bioinformatics Pipeline for Comparative Gene Ontology Analysis in Microbial Genomics

**Authors**

Randy Ortiz1, Torey Looft1

**Affiliations**

1National Animal Disease Center, United States Department of Agriculture, Ames, IA 50010, USA

**Abstract (248 Words)**

Gene Ontology (GO) analysis serves as a foundational tool in bioinformatics for interpreting functional differences among organisms by categorizing genes into standardized terms associated with primary metabolic processes, molecular functions, and cellular components. Despite its utility, comparative GO analyses between groups of microorganisms remain *computationally* *intensive* due to the lack of streamlined, user-friendly tools. Here, we introduce Gene Ontology Overlap Profiler (GOOP), a novel bioinformatics pipeline designed to facilitate the comparison of GO term distributions between two groups of microorganisms. GOOP automates the extraction, normalization, comparison, and visualization of GO terms directly from Bacterial and Viral Bioinformatics Resource Center (BV-BRC) data tables and local genome datasets, providing a cohesive workflow for detecting functional divergences. Key features of GOOP include customizable GO term filtering, integration of comprehensive statistical analyses for identifying significantly enriched or depleted GO categories, and the generation of detailed reports accompanied by an informative heatmap-like graphic. To validate the performance and utility of GOOP, we conducted case studies involving the comparative analysis of two pathogenic bacteria isolated from *Bos taurus*—*Campylobacter jejuni* and *Campylobacter coli*—obtained from the National Center for Biotechnology Information (NCBI) assembly database. Our analyses, conducted at the primary metabolic process level, revealed key differences between the two species which are well documented in scientific literature, as well as differences not well covered in literature. These findings underscore GOOP's potential as an essential tool for microbial genomics, metagenomics, and comparative functional studies, facilitating rapid interpretation of functional differences and uncovering biologically meaningful pathways.

**Introduction (730 Words)**

Understanding the functional differences between microbial species or strains is pivotal in microbial genomics, as it provides insights into pathogenicity, antibiotic resistance, metabolic capabilities, and ecological adaptations. Such knowledge is essential for identifying virulence factors, elucidating mechanisms of antibiotic resistance, and discovering unique metabolic pathways that can inform therapeutic strategies and public health interventions. Two methodologies employed for this purpose are Gene Ontology (GO) analysis and pathway analysis, both of which facilitate the interpretation of genomic data within a functional framework. These analyses are supported by comprehensive databases such as the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), National Center for Biotechnology Information (NCBI), and UniProt, which provide extensive genomic and proteomic data.

GO analysis organizes genes into a structured hierarchy based on three primary domains: primary metabolic processes, molecular functions, and cellular components. This hierarchical framework allows for the classification of genes and their products into categories that reflect their roles within the cell, enabling researchers to identify overrepresented or underrepresented GO terms in a given dataset. By aggregating genes with shared functionalities, GO analysis facilitates the discovery of critical primary metabolic processes and pathways that are active or altered in specific organisms or conditions. For instance, GO analysis can reveal enriched biological processes associated with virulence mechanisms, such as adhesion, invasion, biofilm formation, or toxin production in pathogenic microorganisms.

In contrast, pathway analysis provides a systemic perspective by mapping genes to metabolic or signaling pathways, illustrating how gene products interact within complex biological networks. This approach aids in understanding the regulation of biochemical processes and identifying key nodes or hubs that may be targets for therapeutic intervention. Pathway analysis is particularly valuable for elucidating metabolic capabilities, signaling cascades, and the integration of cellular responses to environmental stimuli.

Despite the utility of GO and pathway analyses, researchers often face challenges in performing comparative functional analyses between groups of microorganisms due to limitations in existing tools and workflows. While resources like the BV-BRC offer extensive data on bacterial pathogens, including genomic sequences, protein structures, and annotated virulence factors, they lack integrated tools for direct GO analysis from data tables available on their platform. Although pathway analysis can be conducted through the BV-BRC's web interface after account creation, the results can be limited by the inherit broadness of pathways which lack the specificity of GO terms. Similarly, databases like NCBI and UniProt provide detailed gene and protein information with functional annotations but do not offer streamlined pipelines for comparative GO analysis across multiple datasets or species.

To address these gaps, we have developed the Gene Ontology Overlap Profiler (GOOP), a novel bioinformatics pipeline designed to automate and simplify the comparative analysis of GO term distributions between two groups of microorganisms. GOOP leverages data from BV-BRC and local genome datasets to extract GO annotations, perform statistical comparisons, and generate comprehensive reports, thereby facilitating the detection of functional divergences that may underlie phenotypic differences or adaptive strategies. Key features of GOOP include customizable GO term filtering, identification of significantly enriched or depleted GO categories, and the generation of detailed GO tables accompanied by a heatmap-like visualization.

In this study, we apply GOOP to genomic datasets from two well documented pathogenic bacteria of significant public health concern: *Campylobacter jejuni* (*C. jejuni*) and *Campylobacter coli* (*C. coli*). *C. jejuni* is a leading cause of bacterial gastroenteritis globally, causing symptoms ranging from diarrhea and abdominal pain to more severe complications like Guillain-Barré syndrome. *C.* *coli*, closely related to *C. jejuni*, also contributes significantly to foodborne illnesses.

Our objective is to utilize GOOP to perform a comparative GO analysis focused on primary metabolic processes among these two well documented pathogens. By starting with BV-BRC data tables, we aim to construct a comparison to identify differences and similarities in their GO terms. This analysis will involve the extraction of GO terms related to carbohydrate primary metabolic processes, normalization of GO term frequencies across datasets, and statistical comparison to identify significant variations. We aim to uncover differences between the organisms that are already well documented in scientific literature as well as eludicate differences that have yet to be explored. Understanding these functional profiles may reveal critical aspects of their pathogenicity, survival strategies, and potential vulnerabilities that can be targeted for control measures or therapeutic interventions. By facilitating a detailed and systematic comparison, GOOP can help uncover subtle yet important functional differences that might be overlooked using traditional analysis methods.

**Methods (1266 Words) <currently WIP>**

In this study, we developed a comprehensive pipeline called Gene Ontology Overlap Profiler (GOOP) to automate the retrieval of genomes, prediction of genes, functional annotation, and comparative analysis of Gene Ontology (GO) terms among microbial datasets. The pipeline integrates multiple bioinformatics tools and databases to facilitate efficient and systematic analysis of genomic data. The process involves several key steps: downloading genomes from the National Center for Biotechnology Information (NCBI) assembly database using data from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), predicting genes with Prodigal, performing functional annotation via BLAST, generating GO count tables, extracting terminal GO terms using the GO directed acyclic graph (DAG), and comparing GO terms between datasets to generate comprehensive reports.

The initial step involves acquiring genomic data from the NCBI using BV-BRC data tables. Using the genome search option in the BV-BRC webpage we filtered by the taxon names *“Campylobacter jejuni”* and *“Campylobacter coli”* with the host name of “cow” and we obtained data tables containing 65 and 19 assembly accessions respectively. Our pipeline utilizes the comma-separated values (CSV) file provided by BV-BRC, which contains a list of genome assemblies with essential columns: “Assembly Accession” and “Genome Name.” These columns serve as identifiers for retrieving the corresponding genome sequences from the NCBI assembly database in which we recovered 63 (2 were duplicates) and 19 genomes for *C. jejuni* and *C. coli* respectively. The pipeline requires several dependencies, including EDirect (NCBI Entrez Direct) for querying NCBI databases, Prodigal for gene prediction, BLASTX for protein sequence alignment, and Python libraries such as argparse, os, subprocess, and pandas for automation.

For each genome listed in the CSV file, we performed automated retrieval and organization using a scripted process. First, the script parses the CSV file to ensure the presence of the necessary columns, “Assembly Accession” and “Genome Name.” An output directory named genomes is created to store the downloaded genome files systematically. The script then iterates over each genome entry, utilizing the esearch and esummary tools from EDirect to query the NCBI assembly database with the “Assembly Accession” number. It retrieves the FTP path for the genome assembly, prioritizing the RefSeq FTP path, and defaults to the GenBank FTP path if the former is unavailable. Using the obtained FTP path, the script constructs the URL to download the genome FASTA file. The wget utility is employed for downloading, and the script includes checks to prevent redundant downloads by verifying the existence of files which in this case two redundant files were detected.

Following genome retrieval, we employed Prodigal, a microbial gene prediction tool, to identify coding sequences within each genome. To ensure compatibility and ease of parsing, the FASTA files were linearized by removing line breaks, and headers were standardized by replacing special characters with underscores. Prodigal was then executed on each genome file using default parameters optimized for prokaryotic genomes. This step provided a set of predicted protein-coding genes for each genome, serving as the basis for subsequent functional annotation.

<<NEED TO ADD SECTION ON GENOMIC COMPLETENESS VALIDATION>>

To assign functional annotations to the predicted genes, we performed BLASTX searches against a custom UniProt protein database. Protein sequences relevant to the taxa under study were retrieved from UniProt based on taxonomy IDs specified (Taxonomy ID 194 for *Campylobacter*) in the input CSV file. These sequences were used to construct a BLAST-formatted protein database using the makeblastdb command, ensuring that the database was comprehensive and tailored to the organisms of interest.

BLASTX was then run for each set of predicted genes against the protein database. The BLASTX searches were configured with an e-value threshold of 1e-10 and at least 80% query coverage to ensure high-confidence matches and included parameters to report the best alignment for each query sequence. This step facilitated the mapping of predicted genes to known proteins, enabling functional characterization based on sequence homology.

With the BLAST results, we proceeded to generate GO count tables, providing a quantitative overview of GO term distributions across the genomes. The BLAST output files were parsed to extract protein accessions, which were then cross-referenced with UniProt GO annotations for *Campylobacter* to retrieve associated GO terms for each gene. The GO terms were counted for each genome, resulting in a summary of GO term frequencies. For each genome, a GO count table was created where rows represented the genome and columns corresponded to unique GO terms. The values within the table indicated the count of each GO term for the respective genome. These tables were compiled into a comprehensive CSV file, enabling comparative analysis of functional profiles among the different microbial genomes.

To focus on primary metabolic processes, we extracted terminal GO terms related to areas of interest using the Gene Ontology DAG. The script accepts a user-provided GO term ID (GO:0044238 for primary metabolic process). It ensures that the latest version of the GO ontology file (go-basic.obo) is available, downloading it if necessary. The GO DAG was loaded using the goatools Python package, which provides structures for traversing the GO hierarchy. We developed a recursive function to build a table of GO terms, capturing the full hierarchy starting from the specified GO term and including all child terms. Terminal GO terms, defined as those without further child nodes, were identified and extracted. These terms represent the most specific annotations within the GO hierarchy and are critical for detailed functional analyses. The terminal GO terms were saved in a tab-separated file with each line containing the GO term ID and its name. This list served as a reference for refining the functional analysis to specific processes, functions, or components relevant to the study's objectives.

The final phase of the pipeline involved comparing GO term distributions between two datasets representing different microbial groups. The previous GO terms list was parsed to extract GO IDs, which were then used to filter and compare GO term counts from the GO count tables of the two datasets. For each GO term in the subset, we compared its counts across species in the query dataset (e.g., one microbial group) against those in the subject dataset (e.g., another microbial group). The differences in counts were calculated, and the maximum difference for each GO term was recorded. GO terms were categorized based on their presence or absence in the datasets producing files for GO terms present in both groups of organisms and GO terms present in either one or the other group only. This comparative analysis enabled the identification of enriched or depleted GO categories between the microbial groups, highlighting functional divergences that may underlie phenotypic differences or adaptive strategies.

<<Include filter for child GO terms>>

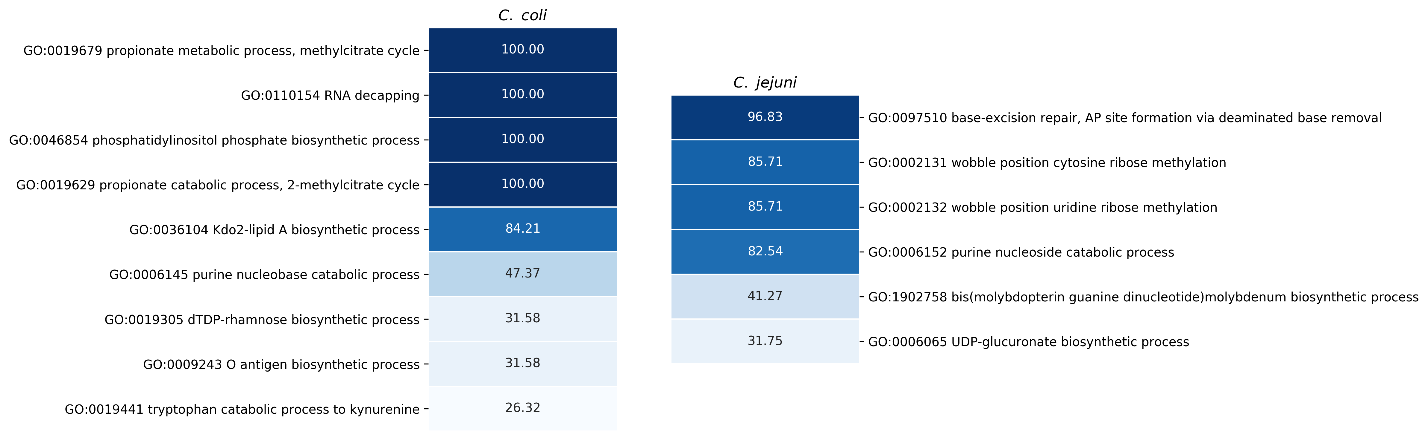
We generated detailed reports summarizing the findings of the comparative analysis. The reports include lists of GO terms unique to each dataset, providing insights into functions specific to each microbial group; counts of GO terms present in both datasets, along with the magnitude of differences, aiding in identifying significant functional variations; and a "max value" block indicating the maximum difference in counts for each GO term across species, highlighting GO terms with the most pronounced differences. To enhance the interpretability of the results, we performed a similarity analysis of the GO term descriptions. Using the Jaccard similarity coefficient, we compared the word content of GO term names (excluding common words) to identify GO terms with similar functional annotations. This analysis provided a nuanced understanding of the relationships between different GO terms and helped identify functionally related categories that may not be collectively depleted. These files provided a comprehensive overview of the functional differences and similarities between the datasets, serving as a valuable resource for further biological interpretation.

**Results (115)**

Comparison of *C. coli* and *C. jejuni* yielded a total of 224 overlapping GO terms and 32 non-overlapping GO terms which were shown to be present in one organism or the other at percentages ranging from complete presence at 100% to partial presence in some of the organisms tested. Of the 34, only 19 terms did not contain a child term that was present in the overlapping GO term group. Groups GO:0046854, GO:0110154, GO:0019629, and GO:0019679 were not only found to be completely missing from *C. jejuni* but present in 100% of *C. coli* organism tested. These results are validated in literature produced by x,x,x,x,x, and x respectively. Four GO groups were filtered due to their presence being less than 15%.

Wagley et al. noted the absence of propanoate metabolism pathway (KEGG: ko00640) which includes GO terms GO:0019629 and GO:0019679.

Wagley, S., Newcombe, J., Laing, E., Yusuf, E., Sambles, C. M., Studholme, D. J., ... & Champion, O. L. (2014). Differences in carbon source utilisation distinguish *Campylobacter jejuni* from *Campylobacter coli*. *BMC microbiology*, *14*, 1-10.



**Figure 1. Gene Ontology Presence / Absence Heatmap for *C. coli* vs. *C. jejuni***

Heatmap of gene ontology terms either present or absent in *C. coli* and *C. jejuni*. Values shown are percentages out of 100. Colors denote proximity to 100% where dark blue is 100% and light blue is proximity to 0%. Values under the *C. coli* title denote GO terms present in *C. coli* but absent in *C. jejuni* (i.e. 100 = missing in 100% of *C. jejuni* datasets but present in at least one *C. coli* dataset) and similarly for the *C. jejuni* column. Values 15% and under were filtered from the diagram.

**Discussion (613)**

Currently there is very little support for gene ontology analyses that work with BV-BRC datasets. We present a comprehensive pipeline capable of comparing two groups of interest against each other to help aid in the field of comparative microbiology.

*C. coli* and *C. jejuni* are two closely related species within the Campylobacter genus, both of which are significant causes of gastroenteritis in humans. Despite their similarities, these species exhibit distinct metabolic capabilities, particularly concerning the propionate metabolic process, the methylcitrate cycle, and lipid A biosynthesis. Understanding these differences involves exploring their genetic makeup, ecological niches, and evolutionary adaptations.

*C. coli* possesses the propionate metabolic pathway, allowing it to utilize propionate—a three-carbon fatty acid—as a carbon and energy source. This capability is particularly advantageous in environments where propionate is available, such as in the gastrointestinal tracts of certain animals or in specific environmental niches. To effectively metabolize propionate, *C. coli* employs the methylcitrate cycle. This metabolic pathway enables the conversion of propionyl-CoA (derived from propionate) into intermediates that can enter the tricarboxylic acid (TCA) cycle, facilitating energy production and biosynthesis. *C. jejuni* does not possess a functional propionate metabolic pathway or the methylcitrate cycle. Instead, it relies primarily on other carbon sources, such as amino acids (e.g., serine and aspartate), for its energy and carbon needs.

The genomes of *C. coli* and *C. jejuni* differ in the presence of specific genes encoding enzymes essential for propionate metabolism and the methylcitrate cycle which is highlighted in our GO analysis. *C. coli* contains genes like prpC (propionyl-CoA carboxylase) and prpD (methylcitrate synthase), which are either absent or non-functional in *C. jejuni*. *C. coli* often inhabits environments where propionate is more readily available, such as the intestines of livestock animals. This ecological context may have driven the retention and optimization of propionate metabolic pathways. Conversely, *C. jejuni* primarily colonizes the avian gut, where different nutrients predominate, reducing the selective pressure to maintain propionate metabolism.

*C. coli* synthesizes lipid A as part of its lipopolysaccharide (LPS) layer, which is crucial for maintaining the integrity of the outer membrane and for interactions with the host immune system. The lipid A biosynthetic pathway in *C. coli* may include specific modifications that confer advantages in certain environments, such as increased resistance to antimicrobial peptides or altered immune recognition. *C. coli* possesses a complete set of genes involved in lipid A biosynthesis, including those encoding enzymes like lpxA, lpxC, and lpxD, which are responsible for the early steps in lipid A assembly. While *C. jejuni* also synthesizes lipid A, the structure and modifications of its lipid A molecules can differ from those of *C. coli*. These differences might result in variations in membrane stability, permeability, and immune evasion strategies.

Such genetic differences can influence how *C. jejuni* interacts with host organisms and responds to environmental stresses. Variations in lipid A structure between *C. coli* and *C. jejuni* can affect how each species is recognized by the host immune system. For instance, differences in lipid A acylation patterns can influence the activation of immune receptors like Toll-like receptor 4 (TLR4), impacting the inflammatory response. Structural differences in lipid A may also contribute to differential resistance to certain antibiotics or antimicrobial agents, as lipid A modifications can alter membrane permeability and stability.

The metabolic distinctions between *C. coli* and *C. jejuni*—specifically regarding propionate metabolism, the methylcitrate cycle, and lipid A biosynthesis—are rooted in their genetic compositions and ecological adaptations. *C. coli* has retained and optimized pathways that allow it to exploit specific carbon sources and survive in particular niches, while *C. jejuni* has evolved alternative strategies suited to its preferred environments.

These differences not only reflect the evolutionary trajectories of these species but also have practical implications for their pathogenicity, interaction with hosts, and responses to antimicrobial treatments. Understanding these metabolic and biochemical distinctions is crucial for developing targeted interventions and treatments for infections caused by these Campylobacter species.

**Conclusion (86 Words)**

The GOOP pipeline offers a robust and automated approach for comparative functional genomics analysis. By integrating genome retrieval, gene prediction, functional annotation, and GO term analysis, it enables researchers to systematically identify and interpret functional differences between microbial groups. The inclusion of terminal GO term extraction and similarity analysis enhances the depth of functional insights, facilitating the discovery of biologically meaningful pathways and processes. This methodology is particularly valuable for studies aiming to understand the functional basis of phenotypic variations, pathogenicity, or environmental adaptations among microorganisms.