Prediction and analysis of Metagenomic operons via MetaRon: a Pipeline for Prediction of Metagenomic OpeRons

Syed Shujaat Ali Zaidi
University of Arizona https://orcid.org/0000-0003-3178-2466
Masood Ur Rehman Kayani
Shanghai Jiao Tong University School of Medicine
Xuegong Zhang
Tsinghua University Department of Automation
Imran Haider Shamsi (drimran@zju.edu.cn)
https://orcid.org/0000-0002-8545-660X

Methodology

Keywords: Escherichia coli, Metagenomic, Operon prediction, Secondary metabolites, Microbiome

Posted Date: February 24th, 2020

DOI: https://doi.org/10.21203/rs.2.24239/v1

Abstract

Background: Efficient regulation of bacterial genes against the environmental stimulus results in unique operonic organizations. Lack of complete reference and functional information makes metagenomic operon prediction challenging and therefore opens new perspectives on the interpretation of the host-microbe interactions.

Methods: Here we present *MetaRon* (pipeline for the prediction of Metagenomic operons), an open-source pipeline explicitly designed for the metagenomic shotgun sequencing data. It recreates the operonic structure without functional information. MetaRon identifies closely packed co-directional gene clusters with a promoter upstream and downstream of the first and last gene, respectively. Promoter prediction marks the transcriptional unit boundary (TUB) of closely packed co-directional gene clusters.

Results: *Escherichia coli* (*E. coli*) K-12 MG1655 presents a gold standard for operon prediction. Therefore, MetaRon was initially implemented on two simulated illumina datasets: (1) *E. coli* MG1655 genome (2) a mixture of *E. coli* MG1655, *Mycobacterium tuberculosis* H37Rv and *Bacillus subtilis* str. 168 genomes. Operons were predicted in the single genome and mixture of genomes with a sensitivity of 97.8% and 93.7%, respectively. In the next phase, operons predicted from *E. coli* c20 draft genome isolated from chicken gut metagenome achieved a sensitivity of 94.1%. Lastly, the application of MetaRon on 145 paired-end gut metagenome samples identified 1,232,407 unique operons.

Conclusion: MetaRon removes two notable limitations of existing methods: (1) dependency on functional information, and (2) liberates the users from enormous metagenomic data management. Current study showed the idea of using operons as subset to represent the whole-metagenome in terms of secondary metabolites and demonstrated its effectiveness in explaining the occurrence of a disease condition. This will significantly reduce the hefty whole-metagenome data to a small more precise data set. Furthermore, metabolic pathways from the operonic sequences were identified in association with the occurrence of type 2 diabetes (T2D). Presumably, this is the first organized effort to predict metagenomic operons and perform a detailed analysis in association with a disease, in this case T2D. The application of MetaRon to metagenome data at diverse scale will be beneficial to understand the gene regulation and therapeutic metagenomics.

Background

Bacteria are one of the simplest free-living forms of life known [1–8]. Due to their presence in diverse environments, these unicellular organisms are susceptible to dynamic conditions [9, 10]. Bacteria can flourish in various conditions through adaptive transcription [11]. Recently, much interest exists in the metagenomics regarding the exploration of novel environments and information such as taxonomic profiling, drug discovery, secondary metabolites and many other [12–16]. Survivalunder various favorable and unfavorable environmental conditions is achieved through the evolution of new proteins, enzymes, and pathways via organizing and clustering of two or more genes into a single structural unit known as an operon [17–22], as shown in Fig. 1. Operon isan organization of genes formed in support of the bacterial system to respond to new environmental stimulus. They are alsovital for the production of natural products many of which have therapeutic importance [23]. This tightly packed co-directional and co-expressed cluster of genes play a crucial role by providing bacterial information about pathways, gene regulation as well as many natural products of industrial importance [24–27].

The influx of information about the uncultured microbes via metagenomics highlighted the importance of metagenomic operons in novel environments. Recent studies have already identified many natural products helpful in treating/prevention of cancer, diabetes, and cholesterol-lowering [28]. Many of which have operonic origins [29, 30]. Operonic insights help in elucidating the diversity and complexity of poorly understood environments. Metagenomic access to novel environments also underscored many essential properties of operons regarding identification of natural products, secondary metabolites, structure and functionality of uncultured microbial communities in association with the disease and environmental conditions [30–33]. Most of the whole-genome operon prediction methods were mostly dependent on experimental or functional information. Such information is often scarce in metagenomic data. Furthermore, the methods developed for whole-genome operon prediction were mostly based on validated information from E. coli and in some instances Bacillus subtilis (B. subtilis). Therefore, predicting operons from a mixture of millions of bacterial species is challenging, but it will also open doors to identify numerous secondary metabolites and the pathways regulating them.

Most of the past metagenomic studies ended exploring the taxonomic classification in context with a particular environment or disease condition, but very few went a step further in analyzing new aspects of metagenomic data [34–38]. Metagenomic operon prediction remains an understudied facet. The scientific community, despite operon's potential contribution, looks forward to a convenient solution; independent of functional and experimental information as well as provides an automated standalone process for metagenomic operon prediction. Furthermore, very few whole-metagenome studies performed a systemic study for the prediction of whole-metagenome operon and analyzing the data from the aspect of operonic secondary metabolites and differentially abundant operonic pathways in association with disease occurrence. To overcome the limitations mentioned above in the prediction of metagenomic operon, we present MetaRon, a Metagenomic operon prediction pipeline for shotgun sequencing data. It is a user-friendly pipeline that performs the necessary downstream data processing (de novo assembly, gene prediction, de novo promoter prediction and gene clustering), before identifying the operons from the metagenomic sample. However, in case of availability of pre-assembled metagenome and genes, MetaRon also predicts the operons, provided the scaffolds are long enough and the format is consistent with the requirements. MetaRon performs operon prediction based on co-directionality, intergenic distance, and promoter parameters.

Operons are clusters of closely packed co-directional genes, highly prevalent in microbes and aidin their survival in diverse conditions. Metagenomic data contains a cumulative mixture of environmental DNA from millions of uncultivable microbes. Operons in these microbes are crucial in understanding the gene regulation, identification of new pathways and discovery of novel products in diverse environmental settings. Identification of metagenomic operons in the laboratory is an intensive and challenging process therefore computational operon prediction is an efficient way to identify operons. However, to our knowledge, no computational pipeline is available to predict metagenomic operons. We have developed MetaRon - a pipeline that performs metagenomic data processing and predicts operons with high sensitivity. This method will be highly beneficial for researchers studying microbial gene regulation, pathways and secondary metabolites.

Methods

Data Sources

MetaRon utilizes multiple data types and sources. Whole-genome of Escherichia coli MG1655 (NCBI RefSeq: NC_000913.3), Bacillus subtilis 168 (NCBI RefSeq: NC_000964), Mycobacterium tuberculosis H37Rv (NCBI Ref Seq: NC_000962) and Escherichia coli C20 draft genome (GenBank accession: NGBR00000000.1) were downloaded from the NCBI, Genome database (https://www.ncbi.nlm.nih.gov/genome). Raw metagenomics reads from the gut of 145 Chinese individuals (Table 1), were retrieved from the European Bioinformatics Institute (EBI), (project ID: SRP008047, https://www.ebi.ac.uk/metagenomics/projects/ SRP008047) [39].

Implementation

MetaRon pipeline is developed and implemented in python 2.7. The application of MetaRon results in several tab delimited and fasta files containing detailed information about predicted genes, operons, gene and operon coordinates, clustering details as well as intermediate processing files that users might be interested in to look at. These outputs will be explained and discussed in detail in this and forthcoming sections.

Table 1
Number of samples belonging to each group of individuals

Category	Count
Disease Lean Female (DLF)	12
Disease Lean Male (DLM)	26
Disease Obese Female (DOM)	13
Disease Obese Male (DOM)	20
Normal Lean Female (NLF)	13
Normal Lean Male (NLM)	24
Normal Obese Female (NOF)	13
Normal Obese Male (NOM)	24

Data Input

MetaRon is a flexible pipeline that accepts input in one of two forms: (1) unassembled data raw reads or (2) metagenomic scaffolds and gene prediction file (.gff file format). MetaRon executes two type of workflows depending on the user input. The process parameter “ago” (Assembly, Gene prediction and Operon prediction) expects trimmed and quality controlled metagenomic reads and performs the required downstream data processing (Fig. 2). This includes de novo assembly via IDBA [40] and prediction of genes via Prodigal [41]. Alternatively, if the user inputs are assembled metagenomic scaffolds and gene prediction file (.gff file format), the user will specify the process parameter “op” (Operon Prediction). The downstream data processing steps of the process “ago” will be skipped in process “op”, and the program starts data extraction on the provided scaffolds and gene prediction files, as shown in Fig. 2.

Feature extraction

Post de novo assembly and gene prediction, the workflow for both “ago” and “op” process is the same. The user also needs to specify the gene prediction tool. MetaRon accepts gene prediction output from Prodigal [41] and MetaGeneMark [42] as legitimate inputs for “op”. However, upon selection of the process “ago”, it will perform gene prediction via Prodigal only (Fig. 2).

The module data_extraction(), mines the gene prediction file (.gff file) according to the specified tool i.e., MetaGeneMark or Prodigal. It parses information such as gene name, gene start, gene end, gene direction, and scaffold name, saving it in a tab-delimited (.tab) file. Next, the module seq_info() extracts information from the scaftig file and creates a dictionary of the scaftigname and length. The output of the data_extraction() and seq_info() is utilized by upstream_coordinates_extraction() and downstream_coordinates_extraction() to calculate the available upstream and downstream region of each co-directional gene clusters. UPS_DSS_Slicing() trims down the upstream and downstream coordinates to 700 bp, if longer. If the upstream or downstream sequences are shorter than 15 bp, it will be assigned a tag “short_ups” and “short_dss”, respectively (Fig. 2 - Feature Extraction). The subsequent step is the extraction of upstream and downstream sequence using the reference coordinates calculated in the previous modules. The getsource function extracts scaftig information from the scaftig file in the form of a dictionary (d) while, the getgenstring_ups(), and getgenstring_dss() modules extracts fasta sequence from the dictionary (d) according to the upstream and downstream coordinates. The upstream fasta sequence is then used to predict the promoters. Upstream or downstream sequences, with the tag “short_ups” and “short_dss”, are ignored in the sequence extrication as well as promoter prediction.

Proximon identification

MetaRon will use the information generated in the previous steps to calculate the intergenic distance (IGD) between co-directional gene clusters via IGD_calc(). Intergenic distance is by far the most common parameter used for the prediction of operons in whole-genomes [19, 25, 27, 43–45]. MetaRon keeps a flexible (< 601 bp) maximum threshold for Intergenic distance. The intergenic distance (IGD) between two genes is calculated as:

IGD (G1,G2)=(start(G2)-end(G1))+1

Where,G1 and G2 are two adjacent co-directional genes, start (G2) refers to the beginning position of second gene in the pair on the genome; while end (G1) refers to the last nucleotide position of the first gene. Co-directional genes with intergenic distance of < 601 bases are clustered as proximons or candidate operons. This threshold is defined as a stretchy parameter due to extremely personalized and diversed definition of IGD in various bacterial species [24]. For the same reason, the proximons identified based on co-directionality and IGD may contain many false positives. The transcription unit boundary could not be accurately defined via the parameters as mentioned above.

Operon prediction

Neural Network Promoter Prediction 2.0 (NNPP) [46] is integrated in the module promoter_prediction(), to predict the upstream promoter for each of the genes in the co-directional closely packed gene clusters. Promoter predictions are parsed and organized via Promoter_file_parse(). Utilizing the outputs of previous modules, Prom_IGD_Clustering() clusters the co-directional genes by intergenic distance and presence of a promoter. At this point, an operon is defined as a cluster of two or more co-directional and closely packed genes with a promoter upstream of the first gene. As shown in Fig. 1, an operon starts with a promoter and ends with a

terminator, however, the presence of the promoter for downstream gene could also signify the end of an operon. Therefore, an operon is a gene cluster delimited by an upstream and downstream promoter signifying the start and end of the operon, respectively.

Unlike Prom_IGD_Clustering(), where co-directionality, IGD and presence of promoter were considered to define an operon, the module Promoter_clustering() predicts the operons with all values of intergenic distances. The pipeline compiles and exports the proximon pairs, and operons in tab-delimited files. Moreover, transitional information such as manipulated gene prediction file, upstream and downstream coordinates and respective sequences, as well as scaftig files are also available to the user for further analysis (Fig. 2 – Operon Prediction).

The implementation of MetaRon on test genomes was followed by application on whole-genome of Escherichia coli MG1655, simulated microbial genomes, Escherichia coli C20 draft genome and lastly on 145 whole-metagenomic samples from the human gut.

Data Analysis

Most of the operon prediction studies focused on efficient operon prediction and little attention to post prediction analysis. We carried out a comprehensive analysis of metagenomic operons, which mainly includes a comparative analysis of biosynthetic gene clusters (BGCs) of operonic sequences and whole-scaffolds as well as the differential pathway analysis of operonic gene clusters. All the 145 gut microbiome samples were divided into eight major group of individuals, based on occurrence of disease, gender and weight (Table 1).

Secondary metabolite identification

Secondary metabolites were identified from operonic sequences and complete scaftigs using anti SMASH (v3.0) (antibiotic and secondary metabolites analysis shell) [47]. The operonic sequences were extracted via operonic cluster coordinate file generated by MetaRon, while complete scaftig sequences were used. Only the top hit proteins per sequence were selected for further analysis. A comparative approach was devised to observe the abundance trend of secondary metabolites in operonic sequences as well as scaftigs for control and type 2 diabetic group of individuals.

Functional mapping and pathway analysis

Raw metagenomic reads were mapped to the operonic sequences using BOWTIE2 [48] and the ".sam" output was processed via SAMtools [49]. Processing includes the conversion of aligned raw read information from sam file format to bam and finally to fastq file format. The raw metagenomic reads aligned to the operonic sequences were further analyzed for differential pathways via a standalone pipeline for functional analysis FMAP [50]. FMAP is a very convenient pipeline that integrates Metagenomic and Metatranscriptomic data and performs differential pathway analysis. Metagenomic raw reads were aligned to the UniRef100 [51] using DIAMOND [52] as the mapping mode. Mapping hits that qualified through the default FMAP settings (sequence identity = > 80%, e-value = > 1e-10) were taken through to the next step of differentially abundant pathways from controls to disease condition. Reads from the previous step were mapped to the KEGG Orthology (KO) database [53, 54]. The mapped reads were normalized to the total number of paired-end reads. The protein gene ID was extracted from the UniProt database [55]. The normalized abundance for each sample was calculated as the number of reads aligned to a gene divided by total read count, followed by a summation of all the genes in the pathway. The pipeline also performs mapping of raw metagenomic reads to the UniRef100 [51] reference

database using DIAMOND [52] and the estimation of gene abundance to identify the differentially abundant pathways and modules.

Results And Discussion

The dependency of previous whole-genome operon prediction methods on experimental and functional information and unavailability of such information in metagenomic data makes metagenonmic operon prediction a tricky task [44, 56–62]. We addressed these limitations via MetaRon, by accurately predicts metagenomic operons without the dependency on functional or experimental information.

MetaRon Implementation

Simulated Genomes

We started with the implementation of MetaRon pipeline on E. coli K-12 MG1655 raw reads simulated via Next-Generation Simulator for Metagenomics (NeSSM) [63]. The microbe is considered as the gold standard for operons. This implementation serves as a litmus test for the performance. The simulation of E. coli genome produced around one million paired end reads of 100 bp length at 20X depth. MetaRon assembled the raw reads via IDBA [64] into 82 scaftigs. The scaftigs with length less than or equal to 500 bp were removed. The remaining scaftigs contains 4,227 genes that were predicted using prodigal [41]. In the first step, MetaRon identified 822 co-directional proximal gene clusters (IGD < 601 bp), containing 2,955 genes. These gene clusters were named as proximons, since they were identified based on direction and intergenic space, as defined by proximon proposition [65–67]. The proximon cluster length range from binary (2 genes) to 32 genes, with no proximons of length 17, 21, 23, 24, 26, 27, 28 and 29 (Fig. 3).

Out of 822 proximons, majority of the clusters (32.9%) are binary while 19.7% and 11.8% proximons are three and four genes long, respectively. The remaining 35.5% of proximons are longer than four genes (Fig. 4). Introduction of a structure defining feature clearly refined the results from proximons to operons. Many genes that that were a part of the proximon cluster are removed by adding the promoter parameter hence, the number of genes in each cluster reduced, leaving behind co-directional closely packed genes that are under the control of a single promoter. This means, an increase in the percentage of binary operons, three and four gene operons but a decrease in clusters with length more than four genes. The proportion of operons with length 2–4 increased to 78% as compared to 64.5% of proximon clusters (Fig. 4). At this point, it is imperative to highlight that no Transcription Unit Boundary (TUB) is definedin the proximal gene clusters. This means that a proximon or a candidate operon might enclosemore than one operon or non-operonic genes. Therefore, we added upstream promoter as a more stringent and structure defining parameter to outline the transcription start and end for an operon (Fig. 1). Prediction of the upstream promoter in the proximal gene clusters results in the removal of non-operonic genes (false positives) and definition of TUBs. This resulted in a total of 828 operons containing 2,893 genes. The longest operon contained 16 genes [68–71]. In comparison with the operons from DOOR database [69, 71], MetaRon achieved a true positive rate of 97.8%. The percentage of binary settings in case of operons increases to 43.9% (364 operons) while the percentageof operon length 3 and 4 is 21.2% (176) and 13.2% (110), respectively (Fig. 4).

These results corroborate with the fact that in E. coli genome, the majority of the operons have binary organization [72, 73]. The percentage of binary gene clusters hold a significant role in accessing the operon predictions since, most of the operons in microbial genomes are binary [27]. An increase in the proportion of such operons in comparison with proximal gene clusters signifies the removal of false positives and improved sensitivity. About 21.7% of operonic clusters have length ranging between five and sixteen (Fig. 4). In order to test its applicability and accuracy, MetaRon was also implemented on raw reads simulated from a whole-genomes of E. coli MG1655, M. tuberculosis H37Rv and B. subtilis 168. The simulation of above mentioned 13,266,813 bp long genomes resulted in two million reads simulated at 15X depth via NeSSM [63]. The resultant 2,514 proximons encompassing 10,625 genes are identified from 232 scaftigs comprising 12,481 genes. The proximons range from 2 to 36 genes in length. In the proceeding step, 2,579 operons containing 8,749 genes are identified. The comparison with the DOOR database demonstrated the sensitivity, specificity, and accuracy of 93.7%, 75.5%, and 88.1%, respectively. Although, MetaRon achieved better performance than previous metagenomic operon prediction work, the performance is affected by reasons such as an operon divided between two scaftigs or the promoter prediction as well as the non-availability of reference. Never the less, the results achieved are encouraging enough to proceed with real data, which is more diverse and complex.

E. coli C20 draft genome operon prediction

We then implemented MetaRon on E. coli C20 draft genome isolated from the metagenome of chicken gut. MetaRon identified 4,544 genes from 4,640,940 bp long genome and resulted in 841 proximons and 946 operons containing 3,937 and 2,409 genes respectively. The longest operon by the number of genes reduced from 33 genes long proximon to 10 genes long operon while the percentage of binary operons significantly increased from 32% (268 proximons) to 71% (673 operons). In this case, MetaRon achieved sensitivity, specificity, and accuracy of 87%, 91%, and 88%, respectively [69, 71]. Majority of operons (68%) discretely mapped to a single reference operon while 20% of predicted operons have over one hit with the reference. Twelve per cent operons demonstrated unique configuration that has less than 50% match with the reference (Fig. 5). This is expected due to the fact that similar genomes could demonstrate variable operonic settings in different conditions [74–77].

Since metagenome data does not have a complete reference on which the raw reads could be mapped, so it is assembled into multiple contigs/scaftigs, rather than in one whole-genome; hencemultiple operonic configurations were observed (Fig. 6). Unlike the proximon proposition, where the majority of the proximons were mapped to more than one operon in a subset fashion, 66% of the operons identified via MetaRon matched precisely to one reference operon as a perfect match. About 8% of the operons show a subset configuration (Fig. 7). A subset configuration refers to an exact match with one or more extra genes in the reference (Fig. 6). While 4% of the predicted operons displayedcontrariwise formation known as a superset configuration, i.e., an operonis longer than the reference operon by one or moregenes (Fig. 7). The subset formations could be due to the distribution of an operon between two scaftigs or different transcription unit boundary (Fig. 6). Furthermore, there were operons that encompasses more than one reference operon in an exact or partial match. Such operonic settings are named as bridge-1, while the other way around is named as bridge-2. About 5% of the above-mentioned bridge settings are observed in the predicted operons. Bridge configurations could be due to altered TUB or the inability of the promoter prediction tool to identify the promoter.

Metagenomic data from various conditions demonstrates new microbial functions under different levels of stress and environmental stimulus [24]. Many unique operonic organizations are likely to appear as a response to new environmental stimuli. This leads to the formation of new or modified operonic configurations such as subsets, supersets or unique operons. In the case of E. coli C20, 17% of predicted operons have less than 50% or no match with the reference (Fig. 7). Such unique organizations may well carry precious insights about the microbial activity for a particular environment regarding bacterial products and pathways [24]. Such insights at metagenomic scale could be valuable in understanding disease condition, its prevention and possibly the cure as well.

Application to Type 2 Diabetes metagenomes

MetaRon was further implemented on whole-metagenomic raw reads from the gut of 145 Chinese individuals (74 Type 2 Diabetic (T2D), 71 controls) [39]. The two groups of individuals are further divided into four sub-groups in each category based on gender, weight and diabetic/non-diabetic (Table 1). MetaRon produced 3,868,389 operons containing 12,414,125 genes (Fig. 8). This makes up almost 50% of the total 23,280,123 genes. There could be similar operons in different samples, so for better organization and an operon catalogue was curated, which resulted in 1.23 million unique operons. The longest operon is 185 genes long. An average 61.3% of the predicted operons showed binary setting. The percentage was consistently high in all samples as shown in Fig. 8.

Prediction of secondary metabolites

Biosynthetic gene clusters (BGCs) were identified from the operons as well as whole-metagenome (Fig. 9). The idea was to explain the occurrence of the disease via secondary metabolites (SMs) and observe the extent of information operons hold in the whole metagenomic assembly. Figure 9 presents a holistic view of the secondary metabolites (SMs) predicted from the operonic sequences and the whole-metagenomes. It can also be observed that there is a notable change in the abundance of SMs from healthy to diabetic state (Fig. 10). Another important observation to highlight is the similarity in the patterns of SMs from operons and whole-metagenomic assembly (Fig. 10). The abundance of the SMs in whole-metagenome was higher than the operonic sequences, which is to be expected, however, the operonic sequences represent nearly the exact pattern as demonstrated via whole metagenomic assembly. The significance of change in abundance of the secondary metabolites from healthy to disease condition was calculated via student's T-test. Several SMs showed significant variance in concentration, as shown Fig. 11.

Functional mapping and analysis

Many functional features of the human gut microbiota have shown correlation with health and disease condition. We evaluate the differential abundance of the operonic pathways in association with health and disease condition. The analysis (See Methodology) was performed between all groups of individuals as mentioned in Table 1. The first analysis was performed on all 145 samples and the results suggest that there were no pathways that demonstrated differential abundance across all control and disease samples. Except one category, i.e., Type 2 Diabetic lean female (DLF) versus healthy lean female (NLF), no variance in patterns was observed across all group of individuals. The result demonstrates a significant downregulation in several pathways from control to the DLF category of the disease group ($p < 0.01$). To validate if the identified pathways

are reported to have association with type 2 diabetes, we tested and found that most of our findings are consistent with the published literature [78–87]. However, here we also report three pathways to have strong association with type 2 diabetes, namely, Maltose phosphorylase (K00691), 3-deoxy-D-glycero-D-galacto-nononate 9-phosphate synthase (K21279) and an uncharacterized protein (K07101). The Maltose phosphorylase catalyzes the phosphorylation process of maltose, resulting in the production of glucose 1-P and glucose. The pathway also overlaps with the glycan degradation [55]. The pathway has never been reported to have any association with T2D, however, glycogen phosphorylase pathway is consistently reported to have strong association with the disease [88, 89]. Further investigation could provide much clear insights into the role of maltose phosphorylase in the occurrence of T2D.

Conclusion

This study presents a convenient publicly available command line pipeline for the processing of Metagenomic data and operon prediction in shotgun sequencing data. A major advantage of MetaRon is that it identifies metagenomic operon without the need of any experimental or functional information, on which most of the previous whole-genome operon prediction methods were based upon. MetaRon is therefore the second pipeline that performs systemic identification of metagenomic operons and the first one to do so, without any prior functional or experimental information. Considering the complexity and incompleteness of metagenomic data, the pipeline predicts metagenomic operons with significantly high specificity. This is the first study to perform a detailed analysis of the metagenomic operons and explaining the occurrence of the disease from the operonic point of view. The analysis reveals a difference in the abundance of secondary metabolites and pathways from the operonic reads, which highlights the role of operons in the metagenomic data. This is also the foremost study to associate the abundance of operonic secondary metabolites with disease and health. Moreover, from the aspect of data management, we demonstrated that operons could also act as a subset to represent the whole-metagenomic sample. As verified through this research, with the extent of information stored in the predicted operons, MetaRon presents many opportunities. MetaRon promises to be a useful pipeline in the identification of metagenomic operons and it is quite certain that more in-depth investigation, aided with wet-lab resources, could provide insightful findings about the diverse microbial biosphere. In this research, the analysis was performed separately on the MetaRon's predicted operons, however, in the future we plan to integrate the prediction of secondary metabolites, pathway annotation and graphical representation within the pipeline.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Please contact author for data requests.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research project is financially supported by the National Basic Research Program of China (2012GB316504), the National Natural Science Foundation of China, International (Regional) Cooperation and Exchange Program, Research fund for International young scientists (517102-N11808ZJ), PSF-NSFC grant 517102-N11909ZJ and Jiangsu Collaborative Innovation Center for Modern Crop Production (JCICMCP) China.

Authors' Contributions

XZ and IHS conceived and designed the study. SSAZ and MRK performed experiments and analyzed the data. SSAZ and MRK contributed to the writing and drafting of the manuscript. XZ and HIS reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgment

The authors acknowledge Dr. Aziz Khan and Kui Hua for insightful discussions regarding this research. Contribution of Ms. Khadija Zahid, Wajeeha Mehdi and Qanita Javed Turabi is extremely valuable in terms of visualization.

References

1. Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. Genome Res. 2009;19:1450–4.
2. Gordo I, Perfeito L, Sousa A. Fitness Effects of Mutations in Bacteria. J Mol Microbiol Biotechnol. 2011;21:20–35.
3. Orr HA, Atwood KC, Schnieder LK, Ryan FJ, Bull J, Badgett M, et al. The distribution of fitness effects among beneficial mutations. Genetics [Internet]. 2003;163:1519–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12702694%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1462510
4. Keightley PD. Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: A simulation study. Genetics. 1998;150:1283–93.
5. Barrick JE, Kauth MR, Strelioff CC, Lenski RE. Escherichia coli rpoB mutants have increased evolvability in proportion to their fitness defects. Mol Biol Evol. 2010;27:1338–47.
6. Mozhayskiy V, Tagkopoulos I. Horizontal gene transfer dynamics and distribution of fitness effects during microbial in silico evolution. BMC Bioinformatics [Internet]. 2012;13:S13. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S10-S13
7. Jiang X, Mu B, Huang Z, Zhang M, Wang X, Tao S. Impacts of mutation effects and population size on mutation rate in asexual populations: a simulation study. BMC Evol Biol [Internet]. 2010;10:298. Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-298

8. Pepperell CS, Casto AM, Kitchen A, Granka JM, Cornejo OE, Holmes EC, et al. The Role of Selection in Shaping Diversity of Natural M. tuberculosis Populations. PLoS Pathog. 2013;9.
9. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. Proc Natl Acad Sci [Internet]. 1998;95:6578–83. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.95.12.6578
10. Torsvik VL, Øvreås L. DNA Reassociation Yields Broad-Scale Information on Metagenome Complexity and Microbial Diversity. Handb Mol Microb Ecol I Metagenomics Complement Approaches. 2011. p. 3–16.
11. Berg JM, Tymoczko JL SL. Prokaryotic DNA-Binding Proteins Bind Specifically to Regulatory Sites in Operons. Biochem 5th Ed. 2002. p. 1282–1284.
12. R L I N C E N T E R F O R G E N O M I C S I N B I O D I V E R S I T Y R E S E A R C H KÖNIGIN-LUISE-S. Recovering genomic clusters of secondary 1 metabolites from lakes: a Metagenomics 2.0 2 approach Background. 2017; Available from: http://dx.doi.org/10.1101/183061
13. Iqbal HA, Low-Beinart L, Obiajulu JU, Brady SF. Natural Product Discovery through Improved Functional Metagenomics in Streptomyces. J Am Chem Soc. 2016;138:9341–4.
14. Gomes ES, Schuch V, de Macedo Lemos EG. Biotechnology of polyketides: new breath of life for the novel antibiotic genetic pathways discovery through metagenomics. Braz J Microbiol [Internet]. 2013;44:1007–34. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3958165&tool=pmcentrez&rendertype=abstract
15. Trindade M, van Zyl LJ, Navarro-Fernández J, Elrazak AA. Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. Front. Microbiol. 2015.
16. Cui H, Li Y, Zhang X. An overview of major metagenomic studies on human microbiomes in health and disease. Quant Biol [Internet]. 2016;1–15. Available from: http://link.springer.com/10.1007/s40484-016-0078-x
17. Rajewsky N. MicroRNAs and the operon paper. J Mol Biol [Internet]. Elsevier B.V.; 2011;409:70–5. Available from: http://dx.doi.org/10.1016/j.jmb.2011.03.021
18. Price MN, Arkin AP, Alm EJ. OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. BMC Bioinformatics. 2006;7:19.
19. Chen X, Su Z, Dam P, Palenik B, Xu Y, Jiang T. Operon prediction by comparative genomics: An application to the Synechococcus sp. WH8102 genome. Nucleic Acids Res. 2004;32:2147–57.
20. Yaniv M. The 50th anniversary of the publication of the operon theory in the journal of molecular biology: Past, present and future. J Mol Biol [Internet]. Elsevier Ltd; 2011;409:1–6. Available from: http://dx.doi.org/10.1016/j.jmb.2011.03.041
21. Jacob F. The birth of the operon. Science. 2011;332:767.
22. Fortino V, Smolander O-P, Auvinen P, Tagliaferri R, Greco D. Transcriptome dynamics-based operon prediction in prokaryotes. BMC Bioinformatics [Internet]. 2014;15:145. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24884724
23. Turnbaugh PJ, Ph D. Moving towards a metagenomic basis of therapeutics.
24. Zaidi SSA, Zhang X. Computational operon prediction in whole-genomes and metagenomes. Brief Funct Genomics [Internet]. 2016;elw034. Available from: https://academic.oup.com/bfg/article-lookup/doi/10.1093/bfgp/elw034

25. Brouwer RWW, Kuipers OP, Van Hijum S a FT. The relative value of operon predictions. Brief Bioinform. 2008;9:367–75.
26. Li G, Che D, Xu Y. A universal operon predictor for prokaryotic genomes. J Bioinform Comput Biol [Internet]. 2009;7:19–38. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19226658%5Cnhttp://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19226658
27. Chuang LY, Chang HW, Tsai JH, Yang CH. Features for computational operon prediction in prokaryotes. Brief Funct Genomics. 2012;11:291–9.
28. Inglis DO, Binkley J, Skrzypek MS, Arnaud MB, Cerqueira GC, Shah P, et al. Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of Aspergillus nidulans, A. fumigatus, A. niger and A. oryzae. BMC Microbiol [Internet]. 2013;13:91. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3689640&tool=pmcentrez&rendertype=abstract
29. Biggins JB, Liu X, Feng Z, Brady SF. Metabolites from the induced expression of cryptic single operons found in the genome of burkholderia pseudomallei. J Am Chem Soc. 2011;133:1638–41.
30. Dumont MG, Radajewski SM, Miguez CB, Mcdonald IR, Murrell JC. Identification of a complete methane monooxygenase operon from soil by combining stable isotope probing and metagenomic analysis. 2006;8:1240–50.
31. Zhang Y, Zhang H. Microbiota associated with type 2 diabetes and its related complications. Food Sci Hum Wellness [Internet]. Beijing Academy of Food Sciences.; 2013;2:167–72. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2213453013000451
32. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature [Internet]. 2013;498:99–103. Available from: http://www.nature.com/doifinder/10.1038/nature12198
33. Nováková J, Farkašovský M. Bioprospecting microbial metagenome for natural products. Biologia (Bratisl) [Internet]. 2013;68:1079–80. Available from: http://www.degruyter.com/view/j/biolog.2013.68.issue-6/s11756-013-0246-7/s11756-013-0246-7.xml
34. Goecks J, Nekrutenko A, Taylor J, Afgan E, Ananda G, Baker D, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11.
35. Seshadri R, Kravitz S a, Smarr L, Gilna P, Frazier M. CAMERA: a community resource for metagenomics. PLoS Biol [Internet]. 2007 [cited 2014 Mar 21];5:e75. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1821059&tool=pmcentrez&rendertype=abstract
36. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One [Internet]. 2012 [cited 2014 Oct 21];7:e47656. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3474746&tool=pmcentrez&rendertype=abstract
37. Markowitz VM, Chen IM a, Chu K, Szeto E, Palaniappan K, Grechkin Y, et al. IMG/M: The integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 2012;40:123–9.
38. Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P. SmashCommunity: A metagenomic annotation and analysis tool. Bioinformatics. 2010;26:2977–8.

39. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature [Internet]. Nature Publishing Group; 2012 [cited 2016 Jun 13];490:55–60. Available from: http://www.nature.com/doifinder/10.1038/nature11450
40. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA - A practical iterative De Bruijn graph De Novo assembler. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2010. p. 426–40.
41. Hyatt D, Chen G, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal : prokaryotic gene recognition and translation initiation site identification. 2010;
42. Ismail WM, Ye Y, Tang H. Gene finding in metatranscriptomic sequences. BMC Bioinformatics [Internet]. BioMed Central Ltd; 2014 [cited 2015 Jan 14];15 Suppl 9:S8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4168707&tool=pmcentrez&rendertype=abstract
43. Moreno-Hagelsieb G. The power of operon rearrangements for predicting functional associations. Comput Struct Biotechnol J [Internet]. The Author; 2015;13:402–6. Available from: http://linkinghub.elsevier.com/retrieve/pii/S200103701500032X
44. Chuang L-Y, Tsai J-H, Yang C-H. Operon Prediction Using Particle Swarm Optimization and Reinforcement Learning. 2010 Int Conf Technol Appl Artif Intell. 2010;366–72.
45. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in Escherichia coli: genomic analyses and predictions. Proc Natl Acad Sci U S A. 2000;97:6652–7.
46. Reese MG. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem. 2001;26:51–6.
47. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 2015;43:W237–43.
48. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
50. Kim J, Kim MS, Koh AY, Xie Y, Zhan X. FMAP : Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. BMC Bioinformatics [Internet]. BMC Bioinformatics; 2016;1–8. Available from: http://dx.doi.org/10.1186/s12859-016-1278-0
51. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics [Internet]. 2014/11/13. Oxford University Press; 2015;31:926–32. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25398609
52. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;12:59. Available from: https://doi.org/10.1038/nmeth.3176
53. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res [Internet]. 2011/11/10. Oxford University Press; 2012;40:D109–14. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22080510
54. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res [Internet]. 2004;32:D277–80. Available from: https://doi.org/10.1093/nar/gkh063
55. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res [Internet]. 2018;47:D506–15. Available from: https://doi.org/10.1093/nar/gky1049

56. Price MN, Huang KH, Alm EJ, Arkin AP. A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. 2005;33:880–92.
57. Chen X, Su Z, Xu Y, Jiang T. Computational Prediction of Operons in Synechococcus sp . WH8102. 15:211–22.
58. Bergman NH, Passalacqua KD, Hanna PC, Qin ZS. Operon prediction for sequenced bacterial genomes without experimental information. Appl Environ Microbiol. 2007;73:846–54.
59. Chuang L, Yang C, Tsai J, Yang C. Operon Prediction Using Chaos Embedded Particle Swarm Optimization. 2013;10:1299–309.
60. Edwards MT, Rison SCG, Stoker NG, Wernisch L. A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. Nucleic Acids Res. 2005;33:3253–62.
61. Tran TT, Dam P, Su Z, Poole FL, Adams MWW, Zhou GT, et al. Operon prediction in Pyrococcus furiosus. Nucleic Acids Res. 2007;35:11–20.
62. Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res. 2010;38.
63. Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. NeSSM: A Next-Generation Sequencing Simulator for Metagenomics. PLoS One. 2013;8.
64. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
65. Vey G, Charles TC. MetaProx: the database of metagenomic proximons. Database [Internet]. 2014;2014:bau097–bau097. Available from: http://database.oxfordjournals.org/cgi/doi/10.1093/database/bau097
66. Vey G, Charles TC. An analysis of the validity and utility of the proximon proposition. 2012;215–20.
67. Detlev G, Vey A. The Proximon : Representation , Evaluation , and Applications of Metagenomic Functional Interactions by.
68. Salgado H, Gama-Castro S, Peralta-Gil M, Díaz-Peredo E, Sánchez-Solano F, Santos-Zavaleta A, et al. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. 2006;34:D394–7.
69. Mao F, Dam P, Chou J, Olman V, Xu Y. DOOR: A database for prokaryotic operons. Nucleic Acids Res. 2009;37:459–63.
70. Dam P, Olman V, Harris K, Su Z, Xu Y. Operon prediction using both genome-specific and general genomic information. Nucleic Acids Res. 2007;35:288–98.
71. Mao X, Ma Q, Zhou C, Chen X, Zhang H, Yang J, et al. DOOR 2.0: Presenting operons and their functions through dynamic and integrated views. Nucleic Acids Res. 2014;42:654–9.
72. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. MBio. 2014;5.
73. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S. Computational Identification of Operons in Microbial Genomes Computational Identification of Operons in Microbial Genomes. Genome Res. 2002;1221–30.
74. Bratlie MS, Johansen J, Drabløs F. Relationship between operon preference and functional properties of persistent genes in bacterial genomes. BMC Genomics. 2010;11:71.

75. Price MN, Arkin AP, Alm EJ. The life-cycle of operons. PLoS Genet. 2006;2:0859–73.
76. Nuñez P a., Romero H, Farber MD, Rocha EPC. Natural selection for operons depends on genome size. Genome Biol Evol. 2013;5:2242–54.
77. Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. Nucleic Acids Res. 2001;29:1216–21.
78. Rahman A, Nahar N, Nawani NN, Jass J, Hossain K, Saud ZA, et al. Bioremediation of hexavalent chromium (VI) by a soil-borne bacterium, Enterobacter cloacae B2-DHA. J Environ Sci Heal - Part A Toxic/Hazardous Subst Environ Eng. 2015;50:1136–47.
79. Ptilovanciv EON, Fernandes GS, Teixeira LC, Reis LA, Pessoa EA, Convento MB, et al. Heme oxygenase 1 improves glucoses metabolism and kidney histological alterations in diabetic rats. Diabetol Metab Syndr [Internet]. 2013;5:3. Available from: https://doi.org/10.1186/1758-5996-5-3
80. Chandrakumar L, Bagyánszki M, Szalai Z, Mezei D, Bódi N. Diabetes-Related Induction of the Heme Oxygenase System and Enhanced Colocalization of Heme Oxygenase 1 and 2 with Neuronal Nitric Oxide Synthase in Myenteric Neurons of Different Intestinal Segments. 2017;2017.
81. NAKAJIMA O, SAITOH S, KIMURA T, OSAKI T, VINCENT KP, TAKAHASHI K, et al. Heme Deficiency Causes Impaired Glycogen Synthesis in Skeletal Muscle Leading to Insulin Resistance. Diabetes [Internet]. 2018;67:1716-P. Available from: http://diabetes.diabetesjournals.org/content/67/Supplement_1/1716-P.abstract
82. Simcox JA, Mitchell TC, Gao Y, Just SF, Cooksey R, Cox J, et al. Dietary iron controls circadian hepatic glucose metabolism through heme synthesis. Diabetes [Internet]. 2014/10/14. American Diabetes Association; 2015;64:1108–19. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25315005
83. Wei M, Wang PG. Chapter Two - Desialylation in physiological and pathological processes: New target for diagnostic and therapeutic development. In: Zhang LBT-P in MB and TS, editor. Glycans Glycosaminoglycans as Clin Biomarkers Ther - Part A [Internet]. Academic Press; 2019. p. 25–57. Available from: http://www.sciencedirect.com/science/article/pii/S1877117318301492
84. Wijnhoven TJM, Van Den Hoven MJW, Ding H, Van Kuppevelt TH, Van Der Vlag J, Berden JHM, et al. Heparanase induces a differential loss of heparan sulphate domains in overt diabetic nephropathy. Diabetologia. 2008;
85. Yokoyama H, Sato K, Okudaira M, Morita C, Takahashi C, Suzuki D, et al. Serum and urinary concentrations of heparan sulfate in patients with diabetic nephropathy. Kidney Int. 1999;
86. Lauer ME, Hascall VC, Wang A. Heparan sulfate analysis from diabetic rat glomeruli. J Biol Chem. 2007;
87. Bishop JR, Foley E, Lawrence R, Esko JD. Insulin-dependent diabetes mellitus in mice does not alter liver heparan sulfate. J Biol Chem. 2010;
88. Baker DJ, Timmons JA, Greenhaff PL. Glycogen phosphorylase inhibition in type 2 diabetes therapy: A systematic evaluation of metabolic and functional effects in rat skeletal muscle. Diabetes. 2005;
89. Treadway JL, Mendys P, Hoover DJ. Glycogen phosphorylase inhibitors for treatment of type 2 diabetes mellitus. Expert Opin. Investig. Drugs. 2001.

Figures

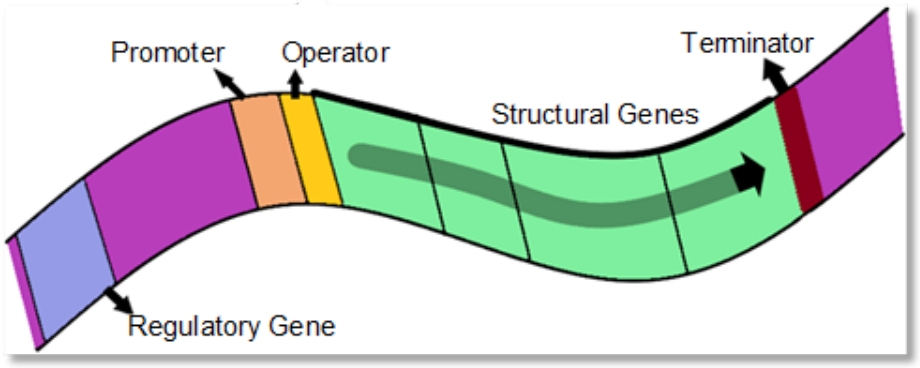

Figure 1

Conceptual structure of an operon.

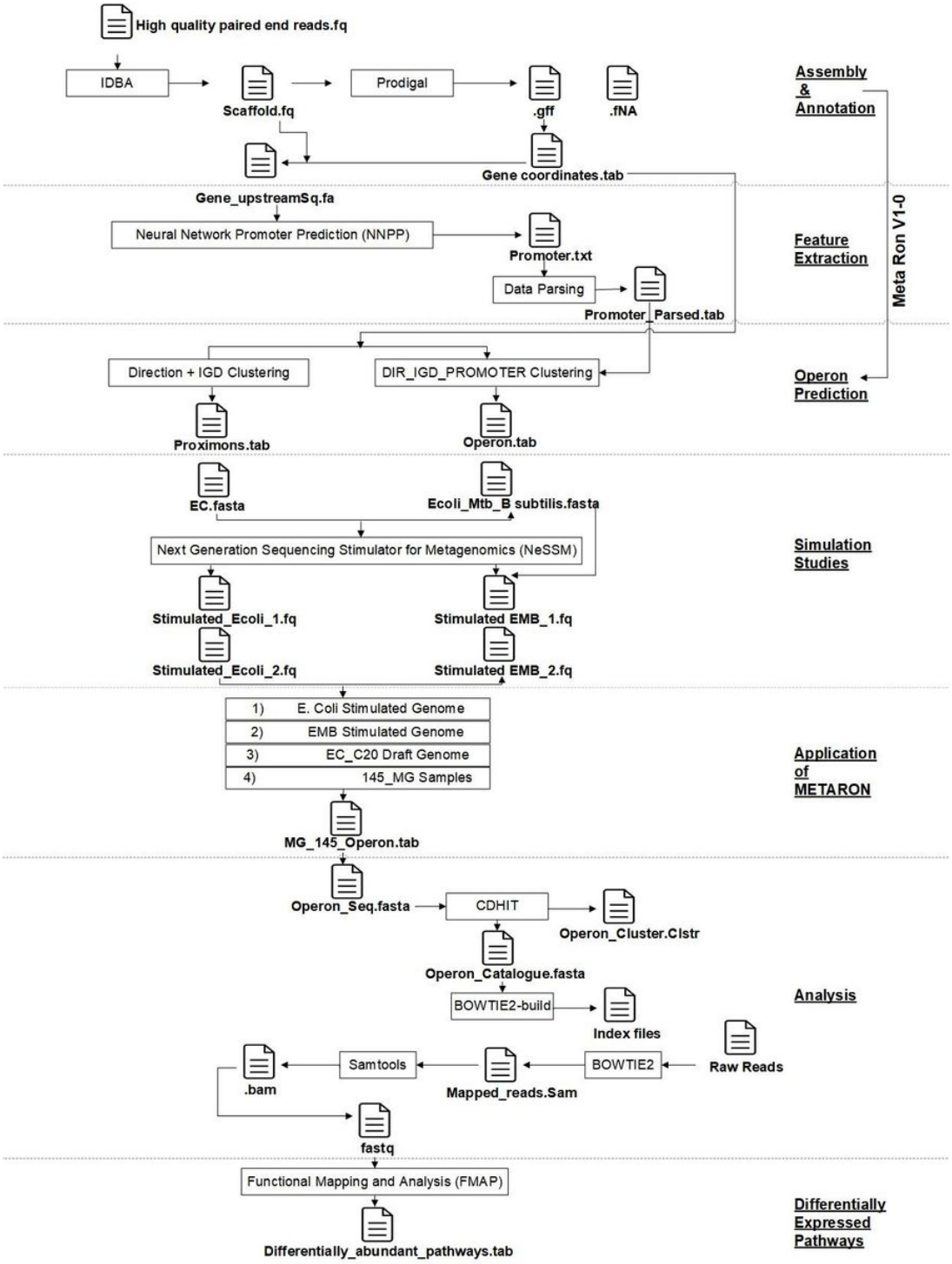

Figure 2

A detailed model demonstrating the prediction and analysis of metagenomic operons via MetaRon.

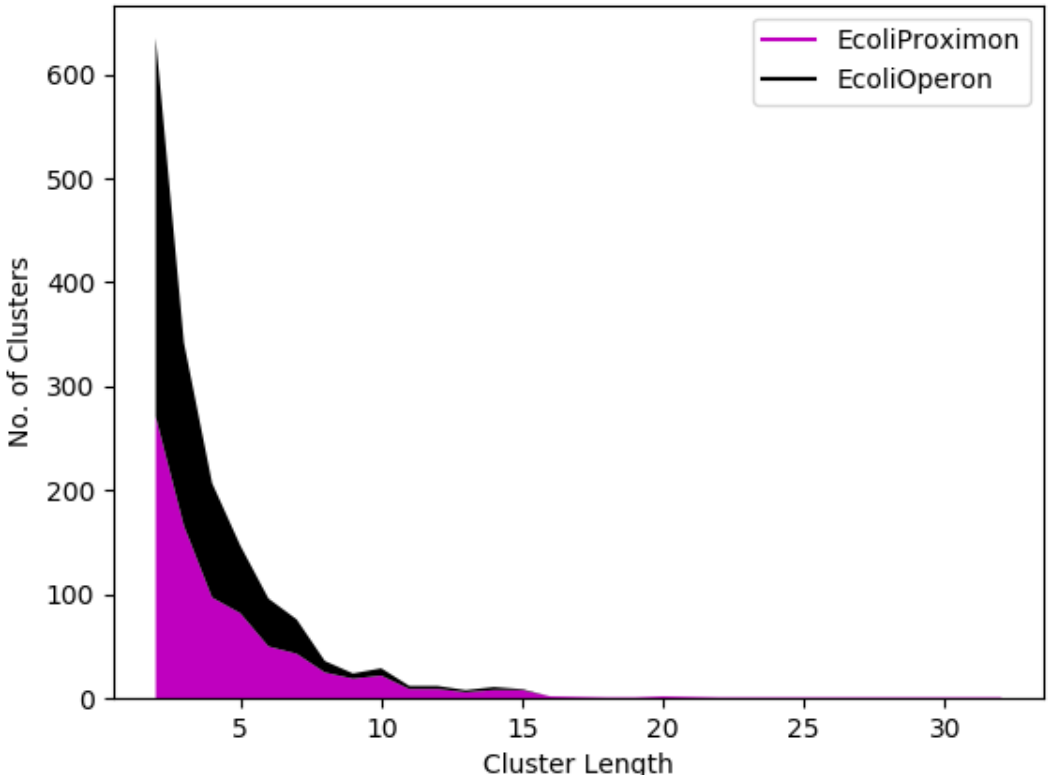

Figure 3

The distribution of operonic and proximonic gene clusters by length.

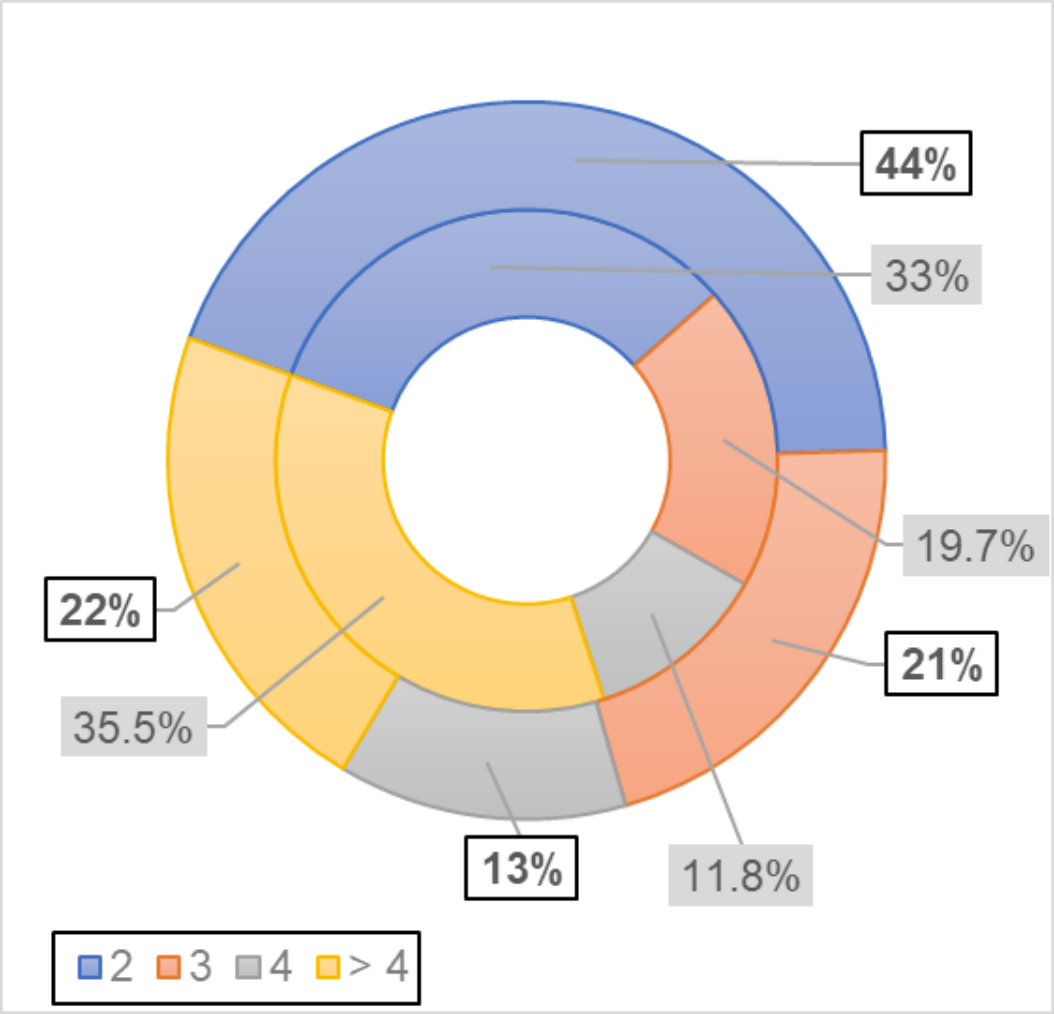

Figure 4

The distribution of operons and proximons in E. coli MG1655.The inner and outer pie chart demonstrates the distribution of proximons and operons by number of genes in each cluster, respectively.

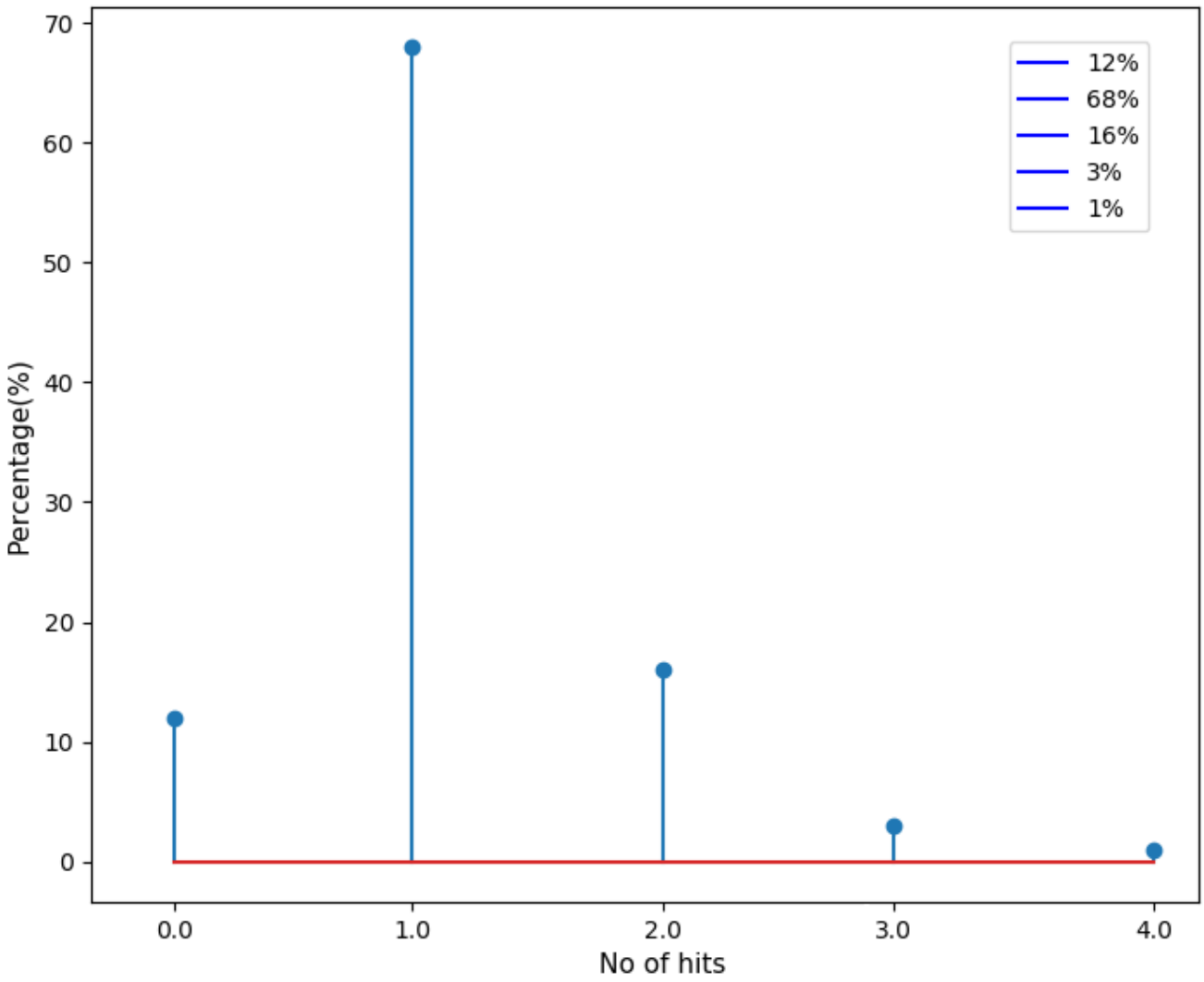

Figure 5

Percentage of E. coli C20 operons mapped to one or more reference operons.

Subset

hycI	hycH	hycG	hycF	hycE	hycD	hycC	hycB	hycA
hycI	hycH	hycG	hycF	hycE	hycD	hycC	hycB	

Superset

patD	ydcS	ydcT	ydcU	ydcV		
ydcS	ydcT	ydcU	ydcV	patD	ydcX	

Bridge-1

	pepP	uniH	ubiI	ygfB	gcvH	gcvP	gcvT
ygfB	pepP	uniH	ubiI	gcvH		gcvP	gcvT

Bridge-2

csiD	lhgO	gabD	gabT	gabP	
csiD	lhgO	gabD	gabT	gabP	csiR

Unique Operonic Organizations

yhjC	yhiD	yhiE

Figure 6

Various operonic configurations from a perfect match to a unique organization.

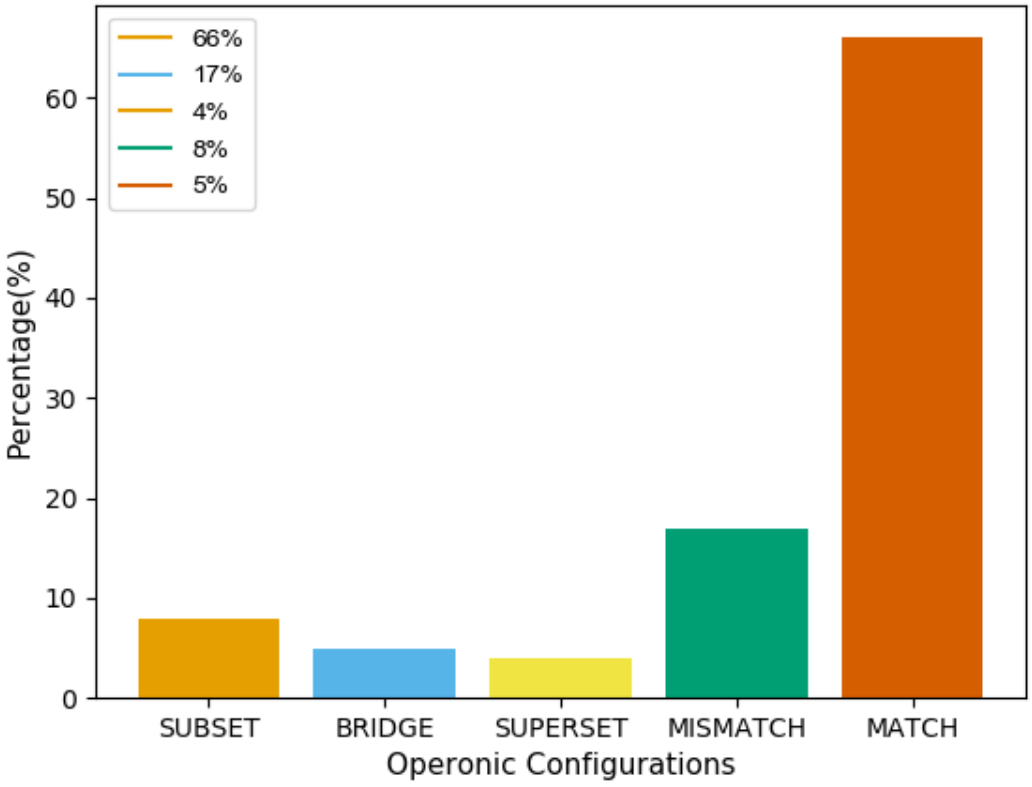

Figure 7

Percentage of operons falling in each operonic category.

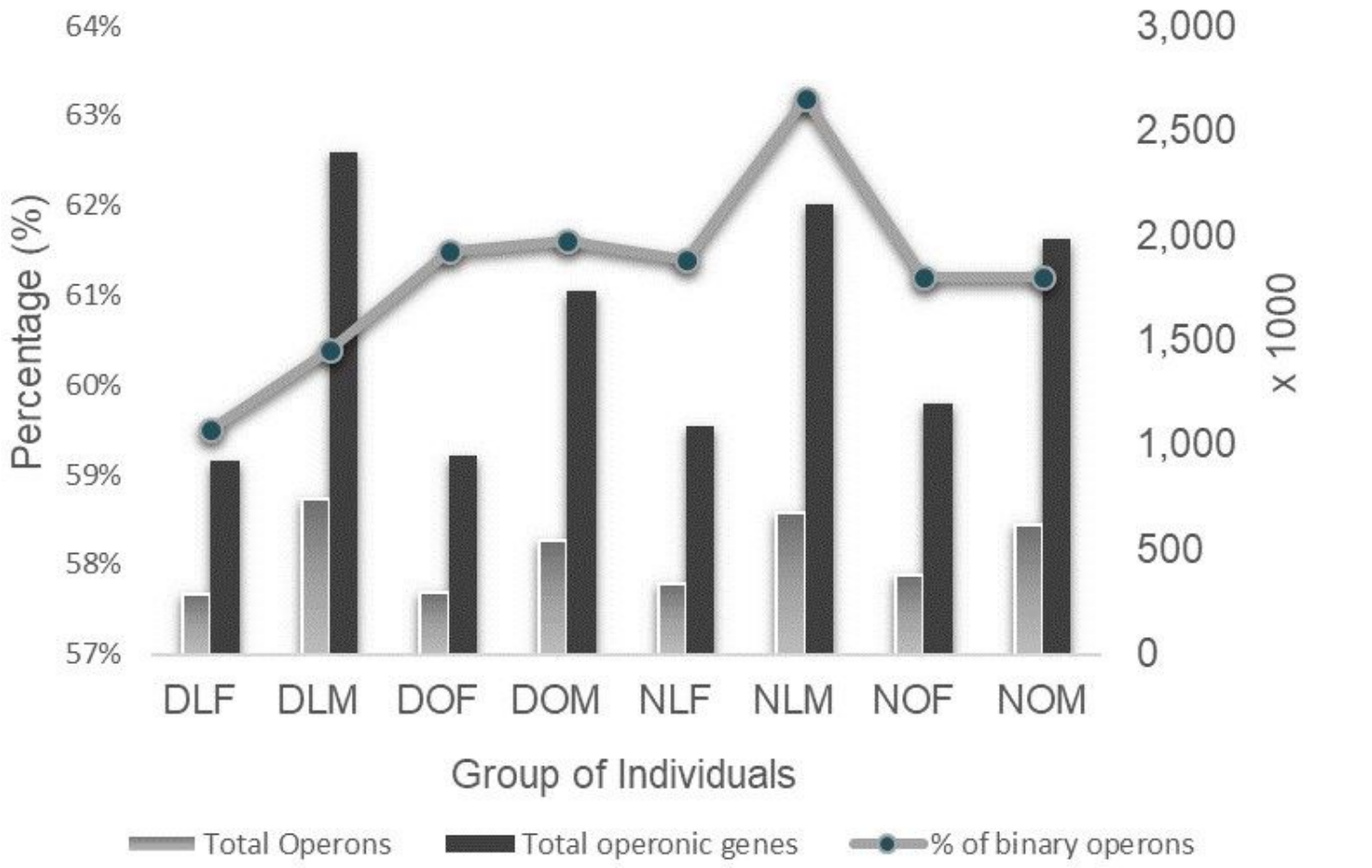

Figure 8

The total number of predicted operons (light grey), the total number of genes present in the operons (dark grey) and the percentage of binary operons.

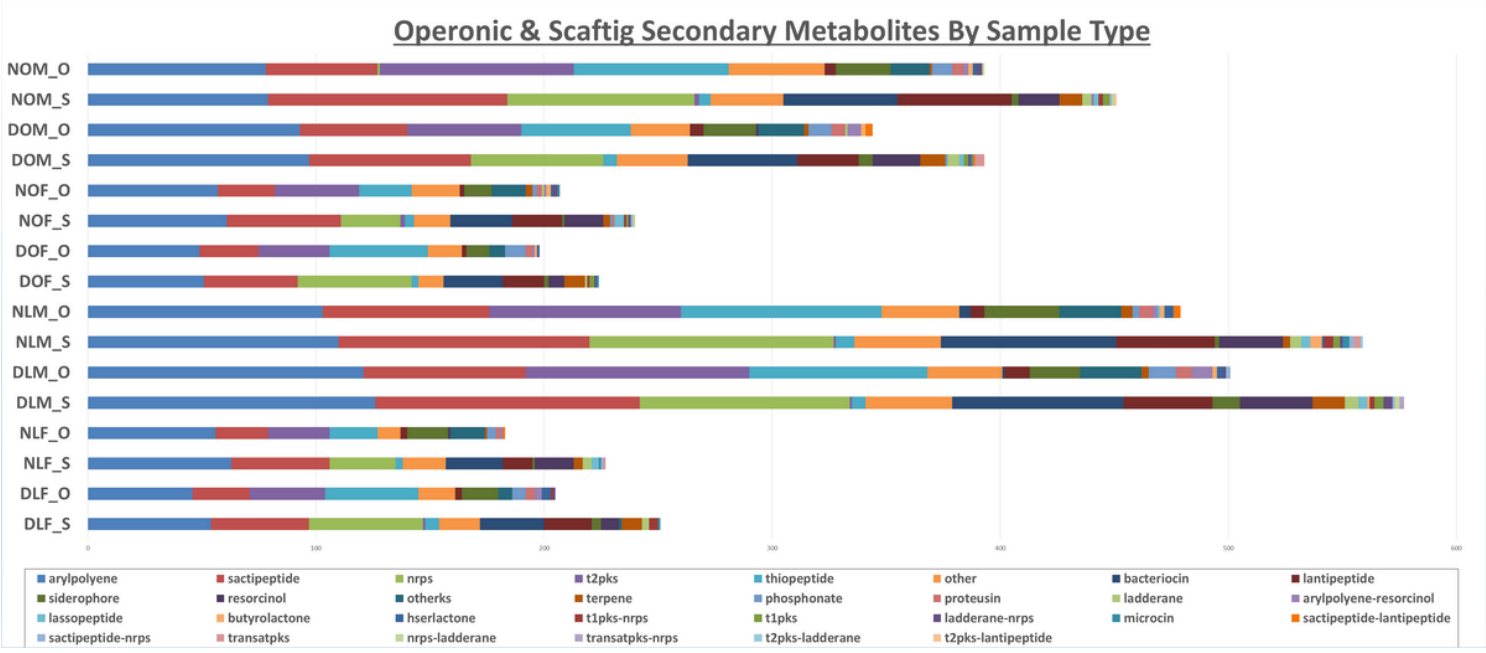

Figure 9

Secondary Metabolites predicted for each group of individuals from Scaftig and operonic sequences.

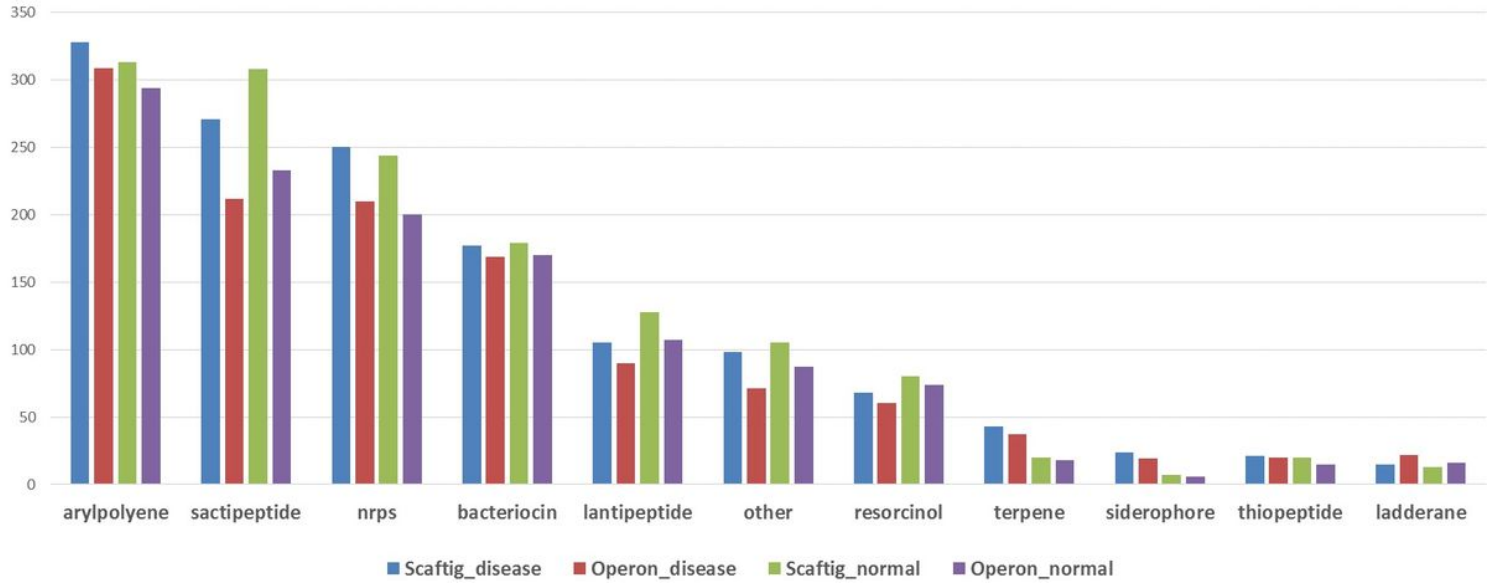

Figure 10

Abundant Secondary Metabolites (SMs) predicted from whole assembly as well as from operonic sequences only.

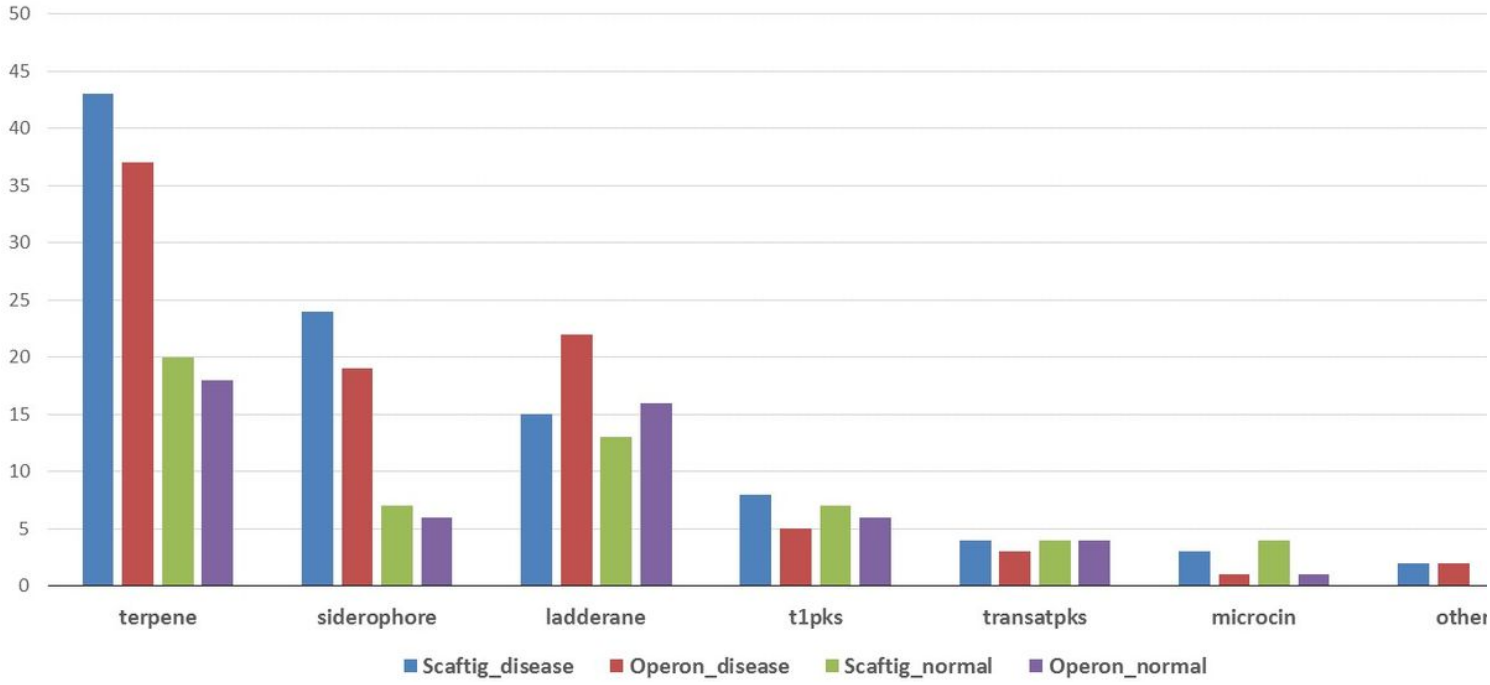

Figure 11

Secondary metabolites significantly abundant differentially.