


SOFTWARE

Open Access



Prediction and analysis of metagenomic operons via MetaRon: a pipeline for prediction of *Metagenome* and whole-genome *opeRons*

Syed Shujaat Ali Zaidi^{1,2,3}, Masood Ur Rehman Kayani⁴, Xuegong Zhang¹, Younan Ouyang⁵ and Imran Haider Shamsi^{6*} 

Abstract

Background: Efficient regulation of bacterial genes in response to the environmental stimulus results in unique gene clusters known as operons. Lack of complete operonic reference and functional information makes the prediction of metagenomic operons a challenging task; thus, opening new perspectives on the interpretation of the host-microbe interactions.

Results: In this work, we identified whole-genome and metagenomic operons via MetaRon (Metagenome and whole-genome opeRon prediction pipeline). MetaRon identifies operons without any experimental or functional information. MetaRon was implemented on datasets with different levels of complexity and information. Starting from its application on whole-genome to simulated mixture of three whole-genomes (*E. coli* MG1655, *Mycobacterium tuberculosis* H37Rv and *Bacillus subtilis* str. 16), *E. coli* c20 draft genome extracted from chicken gut and finally on 145 whole-metagenome data samples from human gut. *MetaRon* consistently achieved high operon prediction sensitivity, specificity and accuracy across *E. coli* whole-genome (97.8, 94.1 and 92.4%), simulated genome (93.7, 75.5 and 88.1%) and *E. coli* c20 (87, 91 and 88%), respectively. Finally, we identified 1,232,407 unique operons from 145 paired-end human gut metagenome samples. We also report strong association of *type 2 diabetes with* Maltose phosphorylase (K00691), 3-deoxy-D-glycero-D-galacto-nononate 9-phosphate synthase (K21279) and an uncharacterized protein (K07101).

(Continued on next page)

* Correspondence: drimran@zju.edu.cn

⁶Department of Agronomy, College of Agriculture and Biotechnology, Key Laboratory of Crop Germplasm Resource, Zhejiang University, Hangzhou 310058, People's Republic of China

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

Conclusion: With *MetaRon*, we were able to remove two notable limitations of existing whole-genome operon prediction methods: (1) generalizability (ability to predict operons in unrelated bacterial genomes), and (2) whole-genome and metagenomic data management. We also demonstrate the use of operons as a subset to represent the trends of secondary metabolites in whole-metagenome data and the role of secondary metabolites in the occurrence of disease condition. Using operonic data from metagenome to study secondary metabolic trends will significantly reduce the data volume to more precise data. Furthermore, the identification of metabolic pathways associated with the occurrence of *type 2 diabetes* (T2D) also presents another dimension of analyzing the human gut metagenome. Presumably, this study is the first organized effort to predict metagenomic operons and perform a detailed analysis in association with a disease, in this case *type 2 diabetes*. The application of *MetaRon* to metagenomic data at diverse scale will be beneficial to understand the gene regulation and therapeutic metagenomics.

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

Background

Bacteria present in diverse environments adaptively transcribe to flourish in dynamic conditions [1–3]. They survive in such conditions through the organization and clustering of two or more genes into a regulatory unit known as an operon [4–9]. Operons play an important role in the evolution of new proteins, enzymes, and pathways; and are vital for the production of natural products - many of which have therapeutic importance [10–14].

Contemporary studies have abundantly identified natural products helpful in treatment/prevention of cancer, diabetes, and lowering cholesterol [15]. Many of these products have operonic origins [16, 17]. Metagenomic access to novel environments also underscored the potential of operons in identification and functionality of uncultured microbial communities (taxonomic profiling, secondary metabolites, drug discovery and many others) [17–25].

Most whole-genome operon prediction methods depend on experimental or functional information in combination with computational parameters [11]; however, experimental/functional information about operons is absent in metagenomic data. Few whole-metagenome studies focused on exploring the operonic aspect of the environment including secondary metabolites and differentially abundant pathways of operonic origin [26–30].

Metagenomic operon prediction thus remains an understudied plane. Operons aiding microbial survival are crucial in understanding the gene regulation, identification of new pathways and novel products in diverse environmental settings. Experimental identification of metagenomic operons is an intensive and challenging process due to everchanging formulation of operons with respect to environmental stimulus. Therefore, computational operon prediction is an efficient way to identify operons. Metagenomic data contains a cumulative mixture of environmental DNA from millions of cultivable and uncultivable microbes. However, to our knowledge, there is no computational pipeline dedicated to predicting metagenomic operons without any functional

information. Considering the importance of operons in bacterial survival, the development of a convenient automated solution independent of functional and experimental information is indispensable.

To overcome the limitations mentioned above, we present *MetaRon*, a *Metagenomic* and whole-genome operon prediction pipeline for shotgun sequencing data. *MetaRon* is a user-friendly pipeline that performs necessary downstream data processing (de novo assembly, gene prediction, de novo promoter prediction and proximon prediction), before identifying the operons from the metagenomic sample. In case of availability of pre-assembled metagenome and genes, *MetaRon* also predicts the operons, directly from scaftigs. The pipeline performs operon prediction with high sensitivity based on co-directionality, intergenic distance, and presence/absence of a promoter upstream and downstream of a gene. This pipeline will be beneficial in studying microbial gene regulation, pathways and secondary metabolites.

Methods

Implementation

MetaRon is developed and implemented in python 3.7. One successful run of *MetaRon* produces several tab delimited and fasta files containing different levels of information. This information will be used for further analysis of metagenomic operons.

Data input

MetaRon executes two type of workflows depending on the user input. The process parameter “**ago**” (Assembly, Gene prediction and Operon prediction) performs downstream data processing using trimmed and quality controlled metagenomic or whole-genome shotgun sequencing reads (Fig. 1). This includes de novo assembly via IDBA [31] and prediction of genes via Prodigal [32]. Alternatively, the user can also input assembled metagenomic scaftigs and gene prediction file (.gff), by specifying the process parameter “**op**” (Operon Prediction).

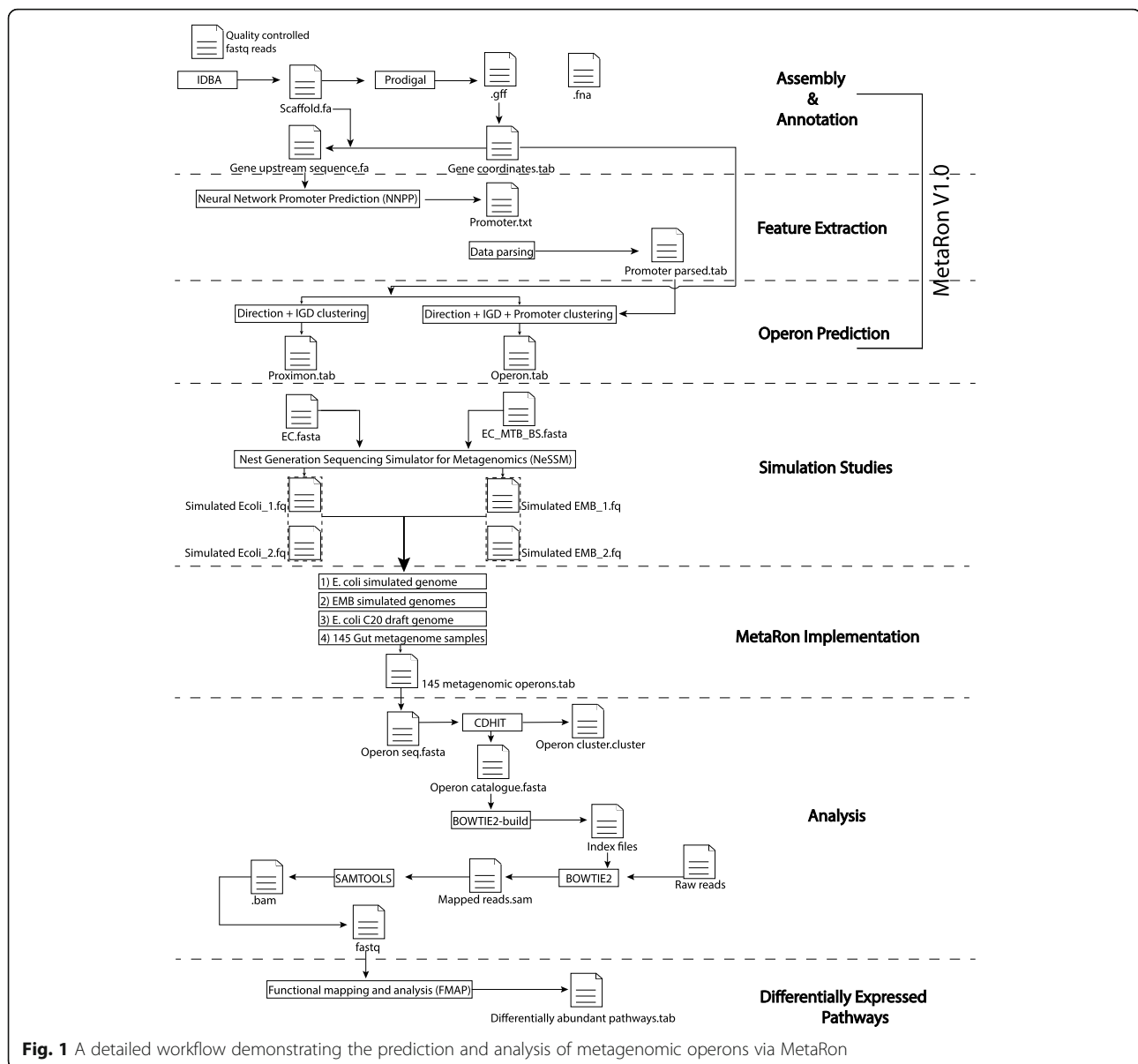


Fig. 1 A detailed workflow demonstrating the prediction and analysis of metagenomic operons via MetaRon

The selection of “op” process will skip the downstream data processing steps directing the program to perform operon prediction only, as shown in Fig. 1. At this point it is important to mention that *MetaRon* only accepts gene prediction files produced by Prodigal and MetaGeneMark. The program requires the user to specify the gene prediction tool used to identify genes.

Feature extraction

Once *MetaRon* reaches the point where it contains de novo assembled scaftigs and gene prediction file, either via process “ago” or “op”, the process of operon prediction is the same (Fig. 1).

1. The *data_extraction()* module mines the gene prediction file (.gff file) and parses information including gene name, gene start and end coordinates, gene direction, and scaftig name into a matrix.
2. Next, the module *seq_info()* creates a dictionary of the scaftig name and scaftig length.
3. The output matrices of *data_extraction()* and *seq_info()* are used to calculate the upstream and downstream intergenic regions of the genes via *upstream_coordinates_extraction()* and *downstream_coordinates_extraction()* modules, respectively.
4. Subsequently, *UPS_DSS_Slicing()* trims down the upstream and downstream coordinates longer than

700 bp to 700 bp. Also, if the upstream or downstream region of a gene is shorter than 15 bp, it will be assigned a tag “short_ups” and “short_dss”, respectively (Fig. 1). These sequences will be ignored in forthcoming steps since signatures for promoter or terminator only appears on/after 15 bp.

- The consequent step is the extraction of upstream and downstream sequence based on the trimmed coordinates (≤ 700 bp). Module *getsource()* extracts scaftig information from the scaftig file in the form of a dictionary (*d*).
- The *getgenstring_ups()*, and *getgenstring_dss()* modules extracts fasta sequence from the dictionary (*d*) using the trimmed upstream and downstream coordinates. The upstream fasta sequence is then used to predict the promoters.

The above-mentioned steps will produce a list of genes with trimmed coordinates and their sequences (upstream and downstream sequences). These coordinates will be used to identify the proximons from the metagenomic data.

Proximon identification

MetaRon will now identify the co-directional gene clusters and calculate the intergenic distance (IGD) (Eq. 1) between the genes in the clusters through *IGD_calc()*. Intergenic distance is by far the most common parameter used for the prediction of operons in whole-genomes [6, 12, 14, 33–35]. The intergenic distance (IGD) between two genes is calculated as:

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

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.

Various operon prediction methods use different range of intergenic distance to identify operons. Based on a thorough review of literature, *MetaRon* defines a flexible (< 601 bp) maximum threshold for Intergenic distance, which was also used by fuzzy genetic algorithm to identify operons [36]. This threshold is defined as a stretchy parameter due to extremely personalized and diverse definition of IGD in various bacterial species [11]. Furthermore, there is no universal threshold for intergenic distance defined for microbes. For metagenomic data, where there are millions of unrelated microbes, a flexible range of intergenic distance will ensure engulfing of all operonic genes in the gene cluster. However, a flexible threshold for intergenic distance will also allow the addition of many non-operonic genes into the cluster. These non-operonic genes will be removed later. Since

these gene clusters are based on proximal genes and co-directionality, they are known as proximons.

The proximons gene clusters also struggle to accurately identify the transcription unit boundary (TUB). Hence, there is a need to accurately identify the transcription unit boundary within each proximons cluster, that will not only remove the non-operonic genes from the cluster but also delimit consecutive operons that were identified as one proximon. These delimited gene clusters with TUB defined will be called operons.

Operon prediction

The module *promoter_prediction()* integrates Neural Network Promoter Prediction 2.0 (NNPP), to predict the upstream promoter for each of the genes in the co-directional closely packed gene clusters [37]. The output is organized into a matrix via *Promoter_file_parse()*. The promoter prediction matrix will be integrated with proximon table and TUBs will be defined, using *Prom_IGD_Clustering()*.

At this moment, 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 the structure of operon indicates, an operon starts with a promoter and ends with a terminator, sandwiching multiple genes within. However, the presence of a promoter downstream of the last gene of the operonic cluster could also signify the end of an operon and start of a new TUB for gene (*i + 1*). Therefore, to redefine, an operon is a gene cluster delimited by an upstream and downstream promoter indicating 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 without considering intergenic distance at all. The pipeline compiles and exports the proximon pairs, and operons in tab-delimited files. Moreover, transitional information such as gene prediction file, upstream and downstream coordinates and fasta files are also available to the user for further analysis (Fig. 1).

MetaRon was implemented on whole-genome, simulated genomes, draft genome and whole-metagenomes, thus demonstrating its performance consistency at different levels of data complexity. The reason was to test the pipeline with different levels of data complexity, both in terms of diversity, information and data format such as, whole-genome or multiple scaftigs. For each of the data input, operons were identified, however, only the metagenomic data was analyzed in detail for its association with *type 2 diabetes* (T2D).

Data analysis

After identification of operons from 145 human gut microbiome samples. We carried out a comprehensive analysis of metagenomic operons, which mainly includes a comparative analysis of biosynthetic gene clusters (BGCs) from operonic origin and whole-scaffig, in addition to the differential pathway analysis from operonic gene clusters.

Secondary metabolite identification

Secondary metabolites were identified separately from operonic and complete scaffig sequences using anti-SMASH (v3.0) (antibiotic and secondary metabolites analysis shell) with default parameters [38]. The operonic sequences were available as the final output file produced by *MetaRon*, while scaffigs were available as the output of de novo assembly in the data processing step of *MetaRon*. A comparative approach was devised to observe the abundance of secondary metabolites in operonic sequences as well as scaffigs for control and *type 2 diabetic* group of individuals.

Functional mapping and pathway analysis

A mapping activity was being carried out all this while where raw metagenomic reads from all 145 samples were individually mapped to the operonic sequences using BOWTIE2 [39]. The resulting 145 sam files were processed using SAMtools [40]. This includes the conversion of sam files to bam and finally to fastq file format. The raw metagenomic reads aligned to the operonic sequences were then analyzed for differential pathways via a standalone pipeline for functional analysis FMAP (Functional Mapping and Analysis Pipeline) [41]. Mapping hits that qualified through the default FMAP settings (sequence identity = > 80%, e-value = > 1e-10) mapped to the KEGG Orthology (KO) database [42, 43]. The mapped reads were then normalized to the total number of paired-end reads. 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. FMAP pipeline also mapped of raw metagenomic reads to the UniRef100 [44] reference database using DIAMOND [45] and estimated the gene abundance to identify the differentially abundant pathways and modules.

Results and discussion

Most of the previous whole-genome operon prediction methods depend highly on experimental and functional information such as microarray data, metabolic pathways, Gene Ontology (GO), and Cluster of Orthologous Groups (COGs). Unavailability of such information in most instances of metagenomic data makes metagenomic operon prediction a tricky task [34, 46–52]. We

addressed these limitations via *MetaRon*, by accurately predicting metagenomic operons independent of functional or experimental information. Although, Vey (2013) demonstrated that metagenomic operons can be identified without any functional or experimental information [53], handling of huge metagenomic data manually is often tedious and prone to errors. Therefore, *MetaRon* presents an automated, improved and universal solution towards the prediction of operons in whole-genome and metagenome shotgun sequencing data.

Data sources

MetaRon utilizes multiple data types and sources. Raw reads of *Escherichia coli* K-12 MG1655 (SRP029211), Whole-genome of *Escherichia coli* MG1655 (NC_000913.3), *Bacillus subtilis* 168 (NC_000964), *Mycobacterium tuberculosis* H37Rv (NC_000962) and *Escherichia coli* C20 draft genome (NGBR00000000.1) were downloaded from the NCBI, Genome database. Human gut metagenomic shotgun sequencing reads from 145 Chinese individuals (Table 1), were retrieved from the European Bioinformatics Institute (SRP008047) [54].

MetaRon application

Whole-genome

E. coli K-12 MG1655 is considered as the gold standard in terms of operons, since it contains the most complete set of operonic information validated experimentally. That is the reason, most of the operon prediction methods were designed and tested on it. We also implemented *MetaRon* on illumine HiSeq reads of *E. coli* K-12 MG1655 as the first run. 82 scaffigs were assembled by *MetaRon* via IDBA [55]. Scaffigs with length less than or equal to 500 bp were removed. The remaining scaffigs resulted in 4227 genes, predicted using prodigal [32]. In the first step, *MetaRon* identified 822 co-directional proximal gene clusters (IGD < 601 bp), containing 2955 genes. These gene clusters were named as proximons, since they were identified based on direction and intergenic space, as defined by proximon proposition [56–

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 |

58]. 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. 2).

Of the 822 proximal clusters, a third of the clusters demonstrated binary configuration, followed by proximons of length three (19.7%), four (11.8) and greater (35.5%). At this point, it is imperative to highlight that no Transcription Unit Boundary (TUB) is defined in the proximal gene clusters. This means that a proximon might enclose more than one operon or non-operonic genes.

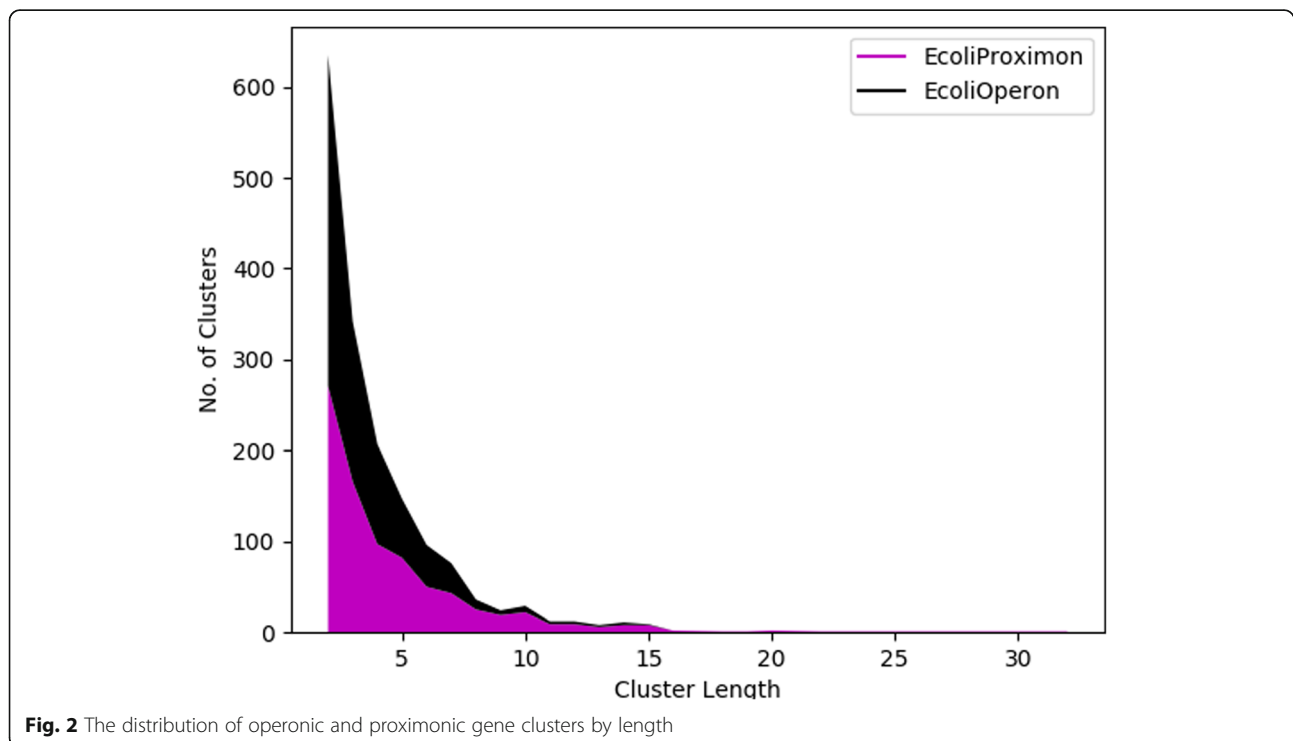
Next, the prediction of promoters further removed the non-operonic genes and clearly defined the transcription unit boundary within the proximons. These filtered proximons are now called operons. The operonic gene clusters contains a promoter upstream of the first and downstream of the last operonic gene. As expected, addition of a stringent structural parameter (promoter) increased the number of operons of length 2,3 and 4 to 364 (43.9%), 176 (21.2%) and 110 (13.2%) operons, respectively. About 21.7% of operons have length ranging between five and sixteen. The proportion of operons with length 2–4 increased to 78% as compared to 64.5% of proximon clusters (Fig. 3). The resultant 828 operons contains 2893 genes while, the longest operon is 16 genes long [59–62]. *MetaRon* achieved a sensitivity, specificity and accuracy of 97.8, 94.1 and 92.4%, respectively, when compared with DOOR database [60, 62].

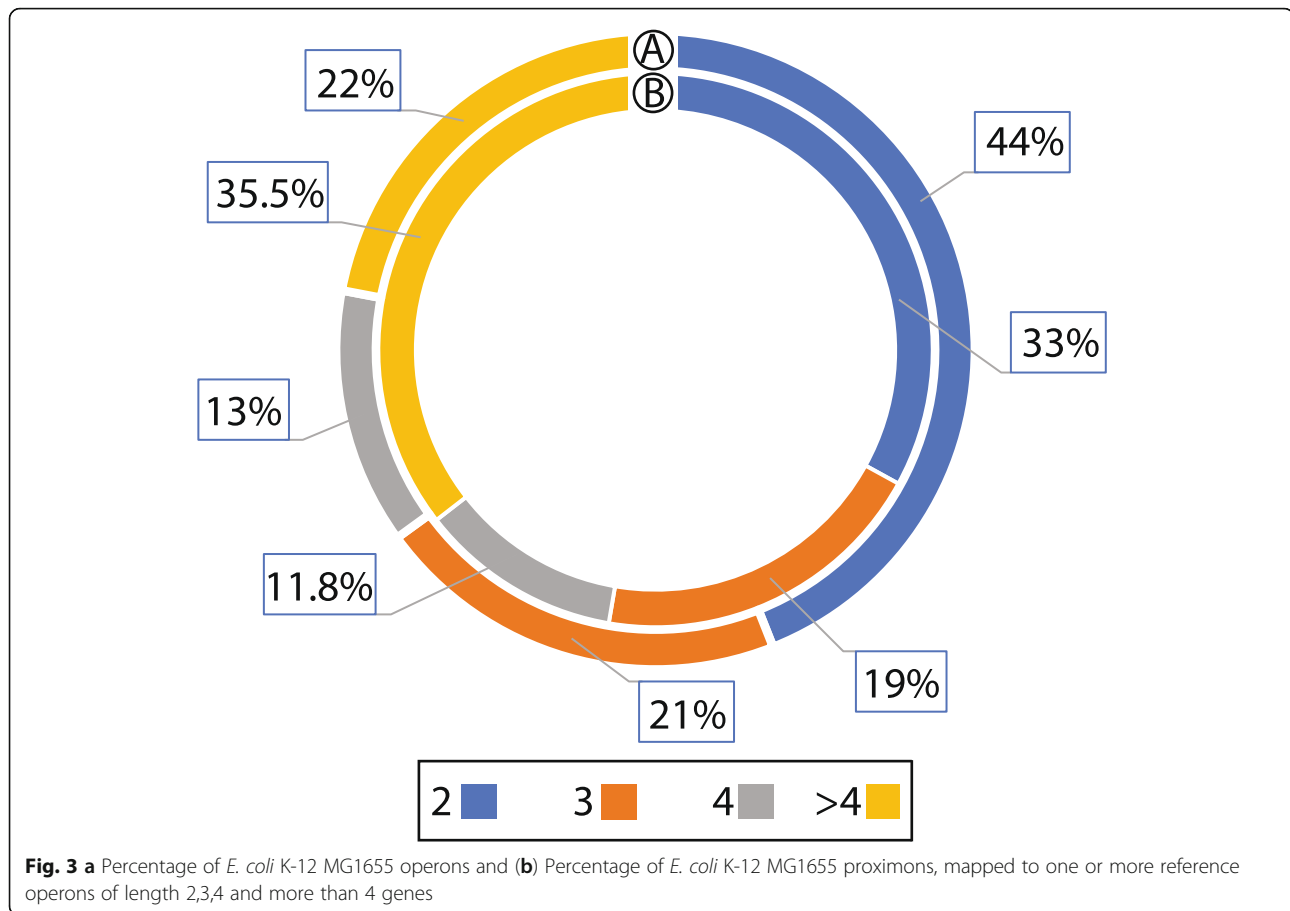
These results corroborate with the fact that most of the operons in *E. coli K12* genome have binary

organization [63, 64]. The percentage of binary operons hold a significant importance in accessing the operon predictions since, most of the operons in microbial genomes are binary [14]. An increase in the proportion of such operons in comparison with proximal gene clusters signifies the removal of false positives and improved sensitivity.

Simulated genomes

In order to test *MetaRon* with more complex data, we simulated illumine raw reads from whole-genomes of *E. coli* MG1655, *M. tuberculosis* H37Rv and *B. subtilis* 168. The sole reason for this simulation was to create a controlled diversity using genomes belonging to the dominant phyla of the microbiome i.e. *B. subtilis* 168 (*firmicutes*), *M. tuberculosis* H37Rv (actinobacteria) and *E. coli* MG1655 (proteobacteria) [65]. The simulation of above mentioned 13,266,813 bp long genomes resulted in two million reads simulated at 15X depth via *NeSSM* (Next-Generation Simulator for Metagenomics) [66]. *MetaRon* assembled the simulated reads into 232 scaftigs containing 12,481 genes. Next, 2514 proximons were identified with a gene count of 10,625 genes. The proximons range from 2 to 36 genes in length. In the proceeding step, 2579 operons containing 8749 genes are identified. On comparison with DOOR database *MetaRon* demonstrated the sensitivity, specificity, and accuracy of 93.7, 75.5, and 88.1%, respectively. Since, there is no metagenomic operon prediction method





available to draw a comparison. We compared *MetaRon* with MetaProx database, which identified proximons and functional gene clusters from the metagenomic data [56]. The results achieved are encouraging enough to move on to more diverse and complex analysis.

E. coli C20 draft genome operon prediction

In the third stage of *MetaRon* implementation and performance evaluation, we identified operons from *E. coli* C20 draft genome isolated from the metagenome of chicken gut. *MetaRon* identified 4544 genes from 4,640,940 bp long genome and resulted in 841 proximons and 946 operons containing 3937 and 2409 genes respectively. The percentage of binary operons significantly increased from 32% (268 proximons) to 71% (673 operons). *MetaRon* achieved a sensitivity, specificity, and accuracy of 87, 91, and 88%, respectively [60, 62].

On comparison with the reference, 68% of the operons discretely mapped to a single reference operon while 20% mapped to more than one operon. Twelve percent of the operons expressed less than 50% identity with the reference hence they were considered as novel or no-hits (Fig. 4). Some variation in the operonic genes could be expected due to the fact that similar genomes could

demonstrate variable operonic settings in different conditions [67–70].

Since metagenome data does not have a complete reference, based on which a reference-based-assembly could be performed, De novo assembly usually produces multiple contigs/scaffolds, rather than one long stretch of DNA; hence multiple operonic configurations were observed (Fig. 5). 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 an exact match with one or more extra gene. This is known as a subset (Fig. 5). 4% of the predicted operons displayed contrary formation known as a superset, i.e., the predicted operon contains one or more extra genes as compared to reference operon. (Fig. 6). The subset formations could be due to the distribution of an operon between two scaffolds or different transcription unit boundary (Fig. 5). Furthermore, there were 5% instances when one predicted operon was matched to more than one consecutive operons (bridge-1) or one reference operon was matched to more than one predicted operon (bridge-2). Bridge configurations could be

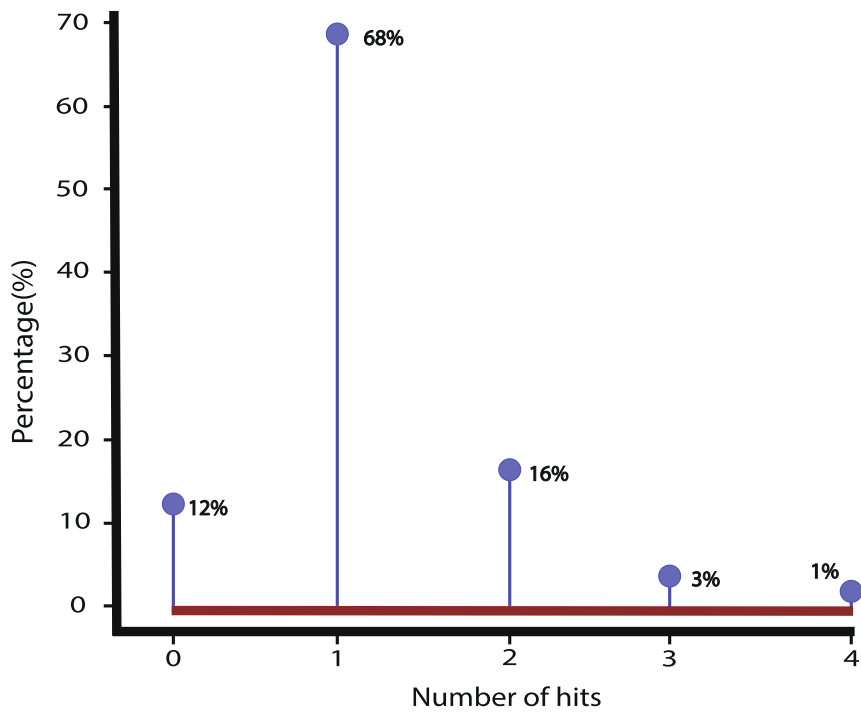


Fig. 4 Distribution of *E. coli* C20 operons by the number of hits, when mapped to the reference genome

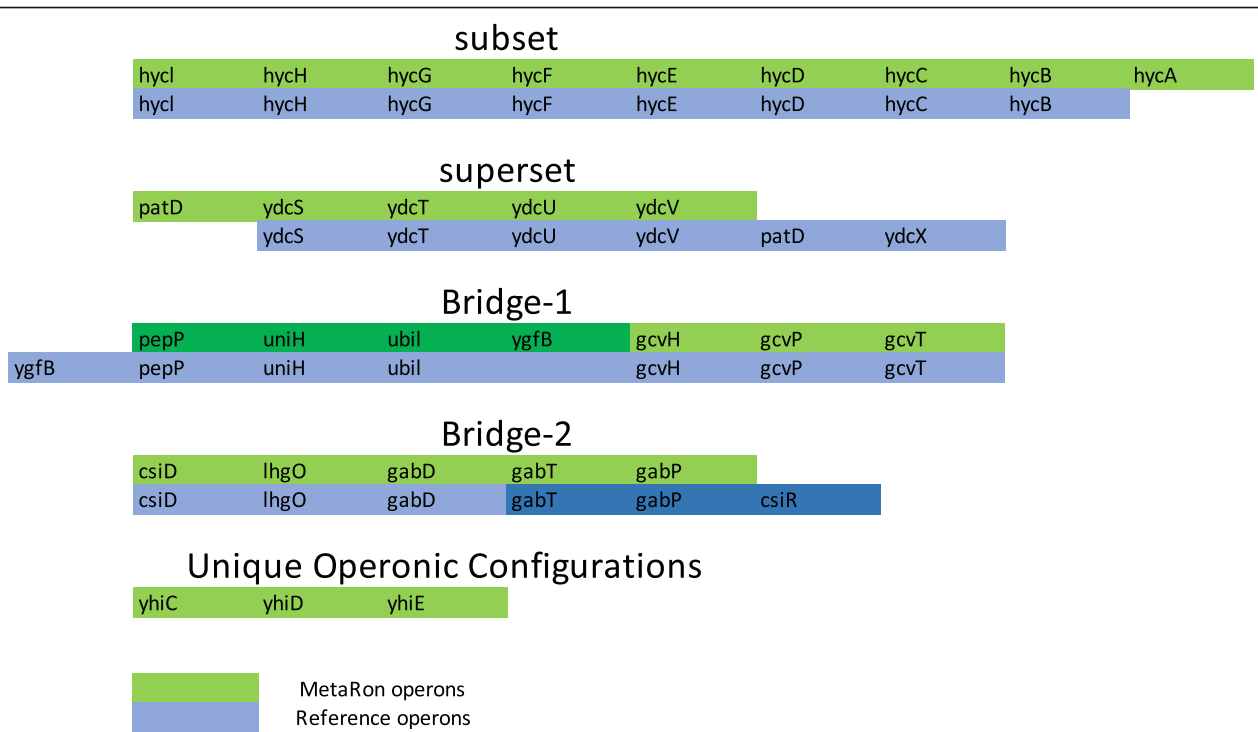
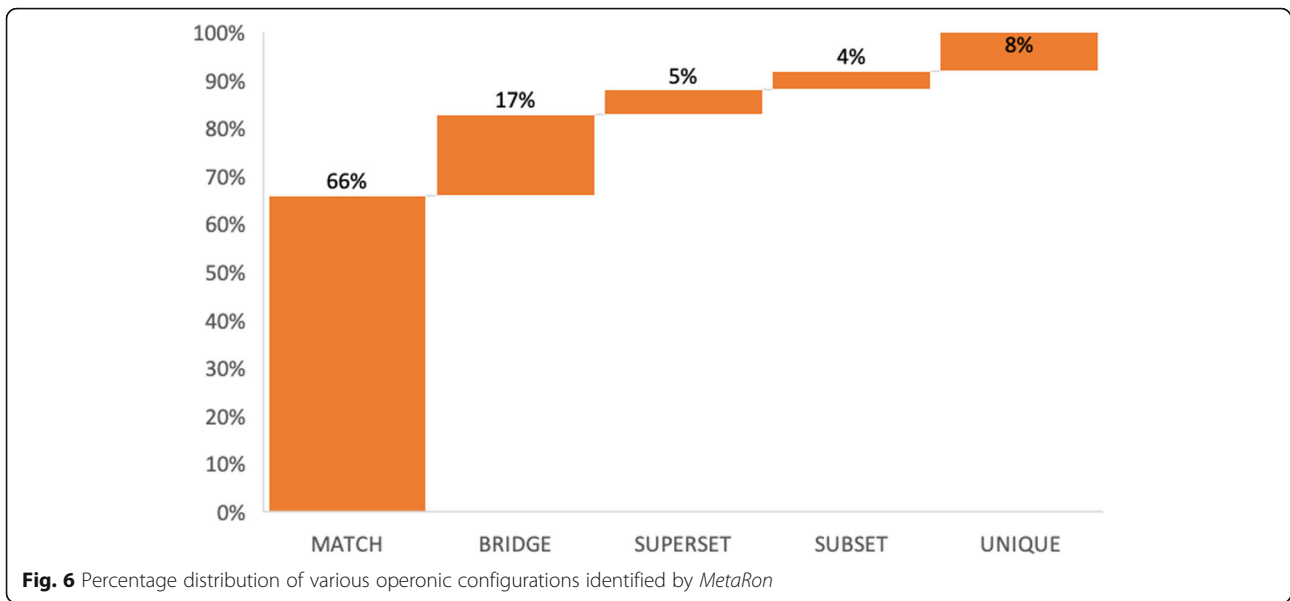


Fig. 5 Operonic configurations observed when operons predicted by *MetaRon* (green) were mapped to the reference operons (blue)



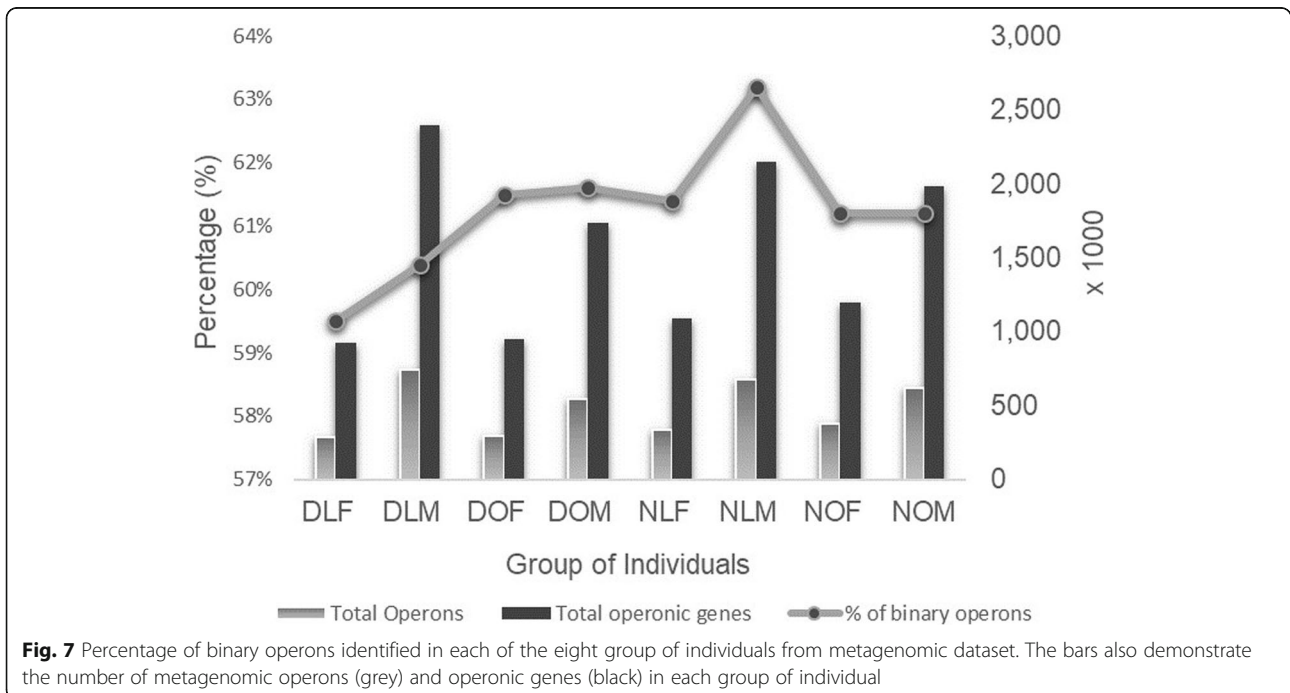
due to altered transcription unit boundary or the inability of the NNPP tool to identify the promoter.

Metagenomic data demonstrates new microbial functions under different levels of stress and environmental stimulus [11]. Many unique operonic organizations are likely to appear as a response to environmental stimuli. This leads to the formation of new or altered 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. 6). Such unique organizations may well carry

precious insights about the microbial activity for a particular environment regarding bacterial products and pathways [11]. 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 shotgun sequencing reads from the gut of 145 Chinese individuals (74 Type 2 Diabetic (T2D), 71 controls) [54]. 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 identified 3,868,389 operons containing 12,414,125 genes (Fig. 7). This makes up almost 50% of the total 23,280,123 genes. Removing operonic redundancy produced 1.23 million unique operons. The longest operon is 185 genes long. The proportion of binary operons was consistently high in all group of individuals (Fig. 7). On average more than 61% operons had binary setting. The non-redundant set of operon sequences will be used for further analysis including identification of biosynthetic gene clusters and differential pathway analysis.

Technical reasons such as quality of assembly and contig/scaffold length could negatively affect the operon prediction. Furthermore, computational promoter prediction being a tough task might result in missing out some operons. Nevertheless, *MetaRon* performed well at all levels of complexity and the above-mentioned reasons would not undermine the utility of *MetaRon*.

Prediction of secondary metabolites

We identified biosynthetic gene clusters (BGCs) from operonic sequences as well as whole-metagenome assembly (Fig. 8). The idea was to demonstrate the

association of disease via secondary metabolites (SMs) and also, observe the extent of information operons hold in the metagenomic data. Figure 8 presents a holistic view of the secondary metabolites (SMs) predicted from the operonic sequences and the metagenomic assembly of each group of individuals. As expected, there is a notable change in the abundance of SMs from healthy condition to diabetic state (Fig. 9). Another novel observation is the similar patterns of SMs in operonic sequences and whole-metagenomic assembly (Fig. 9). We normalized the data to test the significance of change in abundance of the secondary metabolites from healthy to disease condition using student's T-test (95% confidence interval). Several SMs showed significant variance in concentration, as shown in Fig. 10.

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 FMAP analysis (See Methodology) was performed between all groups of individuals as mentioned in Table 1. None of the pathways demonstrated

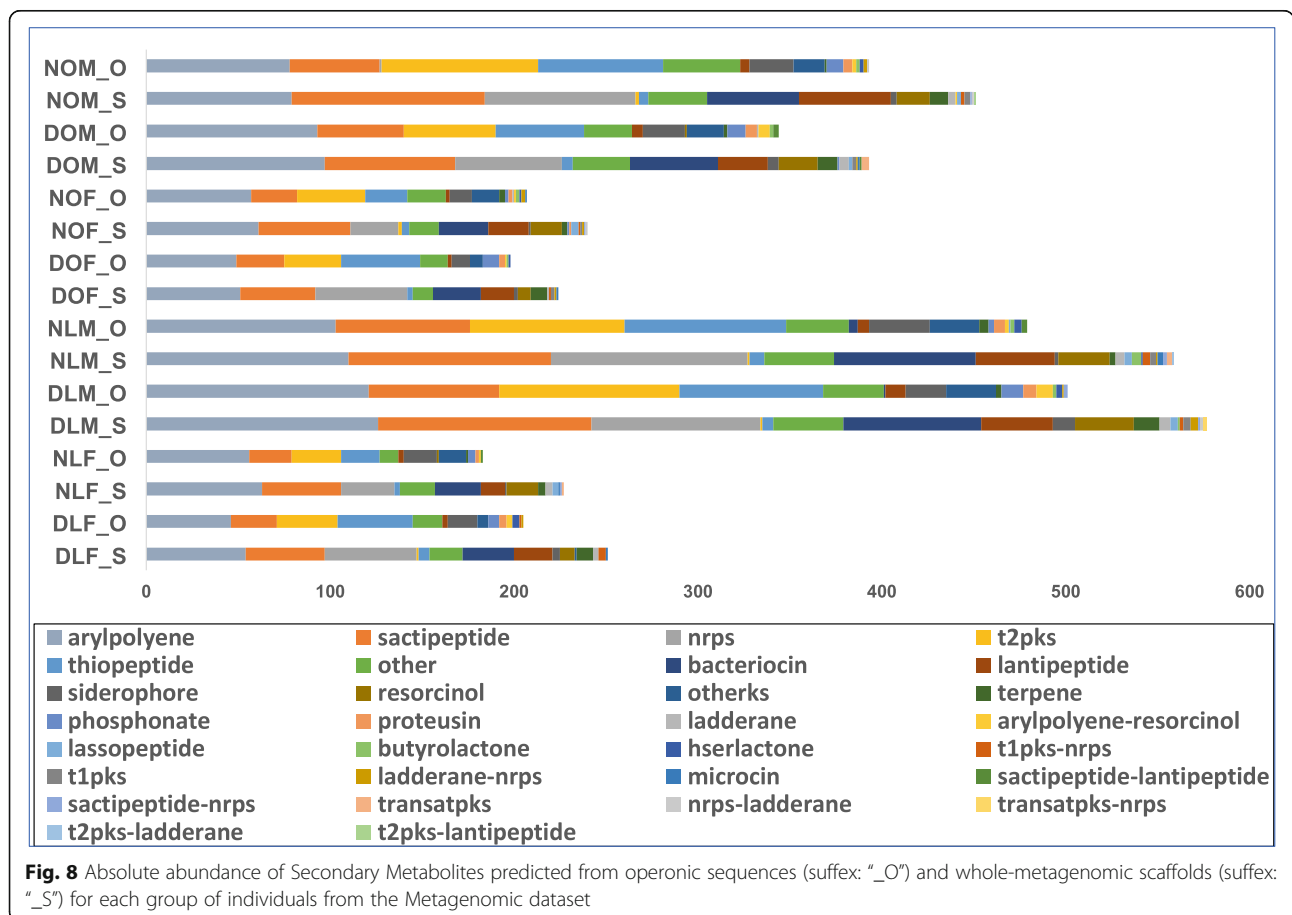
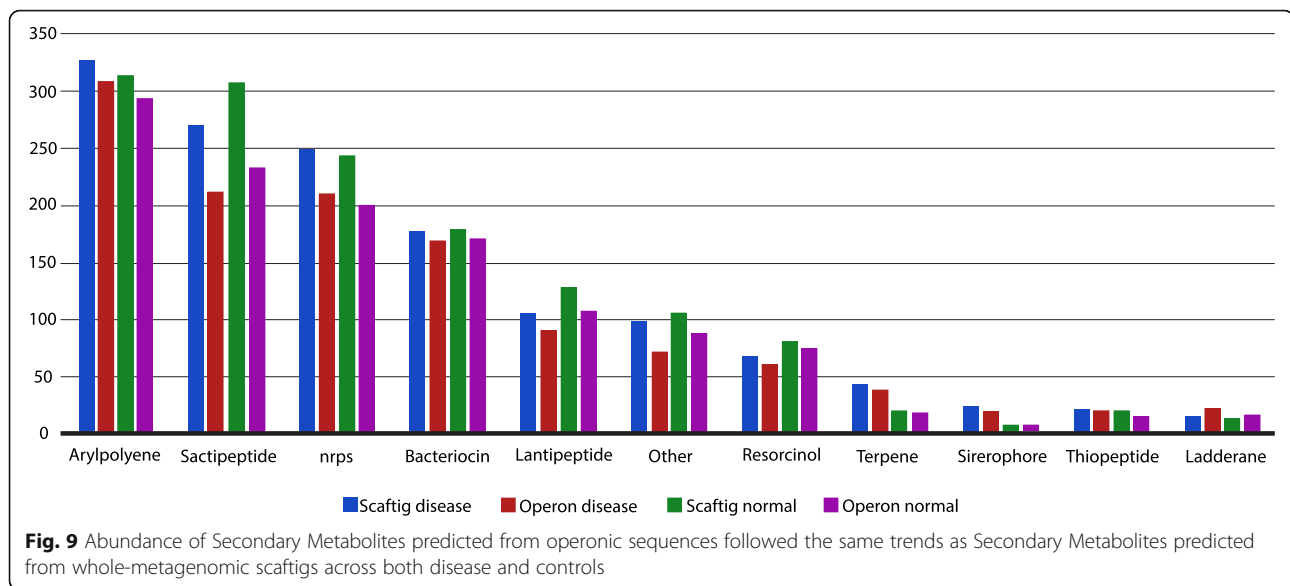
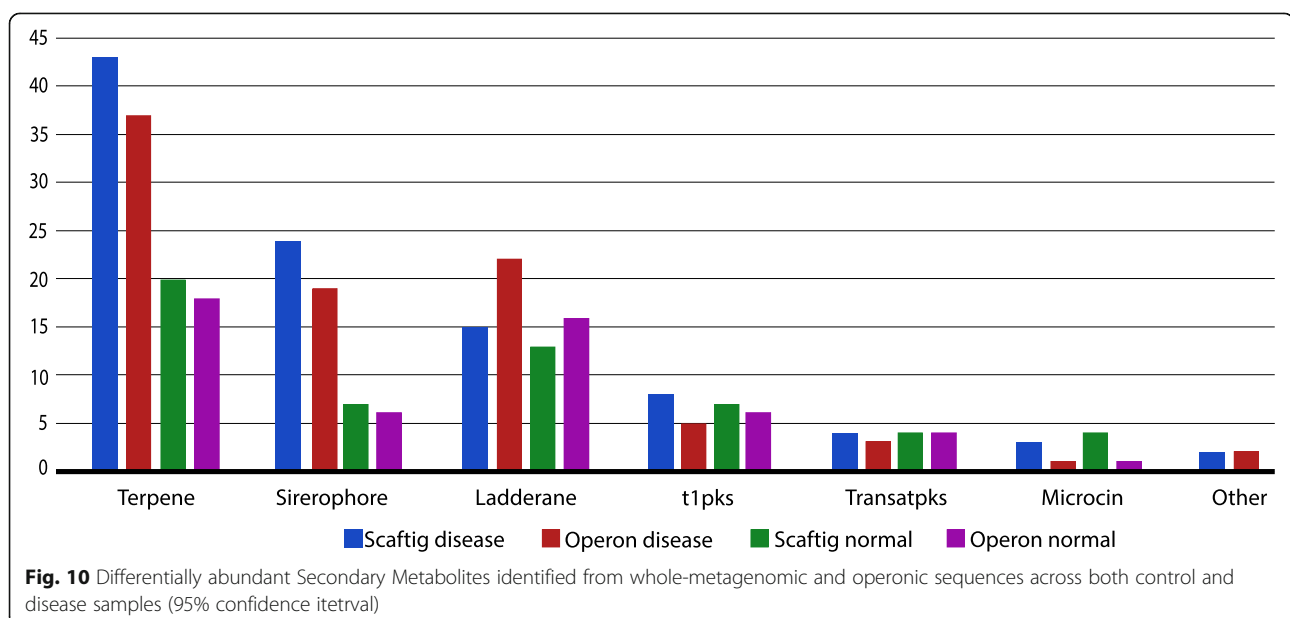


Fig. 8 Absolute abundance of Secondary Metabolites predicted from operonic sequences (suffix: “_O”) and whole-metagenomic scaffolds (suffix: “_S”) for each group of individuals from the Metagenomic dataset



differential abundance across all control and disease samples. With exception of *Type 2 Diabetic* lean female (DLF) versus healthy lean female (NLF). No variance in patterns was observed across any 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 [71–80]. However, here we also report three pathways to have strong association with *type 2 diabetes*, namely, Maltose

phosphorylase (K00691), 3-deoxy-D-glycero-D-galactononate 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 [81]. 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 [82, 83]. 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 independent of any experimental or functional information. *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 very high specificity. This study is also one of the first attempts to perform a detailed downstream analysis of the metagenomic operons and explaining the occurrence of the disease from the operonic point of view.

The differential abundance of operonic secondary metabolites and pathways demonstrated the same trend as of whole metagenome, thus highlighting the amount of information carried by the operons. It also suggests that for the association of secondary metabolites with disease/healthy condition, operons could also act as a subset to represent the whole-metagenomic sample. *MetaRon* promises to be a useful pipeline in the identification of operons from whole-genome and metagenome shotgun sequencing data. 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* predicted operons, however, in the future we plan to integrate the prediction of secondary metabolites, pathway annotation and graphical representation within the pipeline.

Availability and requirements

Project name: *MetaRon* (Metagenomic opeRon Prediction pipeline).

Project Source code availability: <https://github.com/zaidissa/MetaRon>

Operating system: Linux.

Programming language: Python > 3.0.

License: BSD License.

Any restriction to use by non-academics: Academic use only.

Contact: syedzaidi85@hotmail.co.uk, drimran@zju.edu.cn

Acknowledgments

Thanks to Dr. Aziz Khan and Kui Hua for insightful discussions regarding this research and Dr. J. M. Xu for the technical support. Contributions of Ms. Khadija Zahid, Ms. Wajeeha Mehdi and Ms. Qanita Javed Turabi were extremely valuable in terms of visualization.

Authors' contributions

X.Z and I.H.S conceived and designed the study. S.S.A.Z and M.R.K performed experiments and analyzed the data. S.S.A.Z, M.R.K and Y. O contributed to

the writing and drafting of the manuscript. S.S.A.Z, X. Z and I.H.S reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

This research project is financially supported by the National Basic Research Program of China Grant No. 2012GB316504, the National Natural Science Foundation of China, International (Regional) Cooperation and Exchange Program, Research fund for International young scientists Grant No. 31750110462, Sino Pakistan Project NSFC Grant No. 31961143008 and Jiangsu Collaborative Innovation Center for Modern Crop Production (JCIMCP) China.

Availability of data and materials

All data-sets used in the development, testing and analysis of *MetaRon* are publicly available as described in the methods section. Using the tutorial code, user can reproduce the results provided at <https://github.com/zaidissa/MetaRon>. *MetaRon* is designed for Linux. The installation instructions are covered in detail at <https://github.com/zaidissa/MetaRon>

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Bioinformatics Division, Beijing National Research Institute for Information Science and Technology (BNRIST), Department of Automation, Tsinghua University, Beijing 100084, People's Republic of China. ²Bioscience Department, COMSATS Institute of Information Technology, Islamabad 44000, Pakistan. ³Center for Innovation in Brain Science, University of Arizona, Tucson 85719, USA. ⁴Center for Microbiota and Immunological Diseases, Shanghai General Hospital, Shanghai Institute of Immunology, Shanghai Jiao Tong University, School of Medicine, Shanghai 2000025, People's Republic of China. ⁵China National Rice Research Institute (CNRRI), 28 Shuidaosuo rd, Fuyang, Hangzhou 311400, People's Republic of China. ⁶Department of Agronomy, College of Agriculture and Biotechnology, Key Laboratory of Crop Germplasm Resource, Zhejiang University, Hangzhou 310058, People's Republic of China.

Received: 21 April 2020 Accepted: 27 December 2020

Published online: 19 January 2021

References

- Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. *Proc Natl Acad Sci.* 1998;95:6578–83. <https://doi.org/10.1073/pnas.95.12.6578>.
- Torsvik VL, Øvreås L. DNA Reassociation Yields Broad-Scale Information on Metagenome Complexity and Microbial Diversity. In: *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches.* 2011. p. 3–16.
- Berg JM, Tymoczko JL SL. Prokaryotic DNA-Binding Proteins Bind Specifically to Regulatory Sites in Operons. In: *Biochemistry.* 5th edition. 2002. p. 1282–1284.
- Rajewsky N. MicroRNAs and the operon paper. *J Mol Biol.* 2011;409:70–5. <https://doi.org/10.1016/j.jmb.2011.03.021>.
- Price MN, Arkin AP, Alm EJ. OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC Bioinformatics.* 2006;7:19.
- 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.
- 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.* 2011; 409:1–6. <https://doi.org/10.1016/j.jmb.2011.03.041>.
- Jacob F. The birth of the operon. *Science.* 2011;332:767.

9. Fortino V, Smolander O-P, Auvinen P, Tagliaferri R, Greco D. Transcriptome dynamics-based operon prediction in prokaryotes. *BMC Bioinformatics*. 2014;15:145. <https://doi.org/10.1186/1471-2105-15-145>.
10. Turnbaugh PJ, Ph D. Moving towards a metagenomic basis of therapeutics. 2013.
11. SSA Z, Zhang X. Computational operon prediction in whole-genomes and metagenomes. *Brief Funct Genomics*. 2016;elw034. <https://doi.org/10.1093/bfgp/elw034>.
12. Brouwer RWW, Kuipers OP, Van Hijum SA. The relative value of operon predictions. *Brief Bioinform*. 2008;9:367–75.
13. Li G, Che D, Xu Y. A universal operon predictor for prokaryotic genomes. *J Bioinform Comput Biol*. 2009;7:19–38 doi: S0219720009003984 [pii].
14. Chuang LY, Chang HW, Tsai JH, Yang CH. Features for computational operon prediction in prokaryotes. *Brief Funct Genomics*. 2012;11:291–9.
15. 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*. 2013;13:91. <https://doi.org/10.1186/1471-2180-13-91>.
16. 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.
17. 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. *Environ Microbiol*. 2006;8(7): 1240–50. <https://doi.org/10.1111/j.1462-2920.2006.01018>.
18. Cuadrat RRC, Ionescu D, Dávila AMR, Grossart HP. Recovering Genomics Clusters of Secondary Metabolites from Lakes Using Genome-Resolved Metagenomics. *Front Microbiol*. 2018; 20:9:251. <https://doi.org/10.3389/fmicb.2018.00251>.
19. 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.
20. 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*. 2013;44:1007–34 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3958165&tool=pmcentrez&rendertype=abstract>.
21. Trindade M, van Zyl LJ, Navarro-Fernández J, Abd Elrazak A. Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Front Microbiol*. 2015;28(6): 890. <https://doi.org/10.3389/fmicb.2015.00890>.
22. Cui H, Li Y, Zhang X. An overview of major metagenomic studies on human microbiomes in health and disease. *Quant Biol*. 2016;1–15. <https://doi.org/10.1007/s40484-016-0078-x>.
23. Zhang Y, Zhang H. Microbiota associated with type 2 diabetes and its related complications. *Food Sci Hum Wellness*. 2013;2:167–72. <https://doi.org/10.1016/j.fshw.2013.09.002>.
24. 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*. 2013;498:99–103. <https://doi.org/10.1038/nature12198>.
25. Nováková J, Farkašová M. Bioprospecting microbial metagenome for natural products. *Biologia (Bratisl)*. 2013;68:1079–80. <https://doi.org/10.2478/s11756-013-0246-7>.
26. 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.
27. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. CAMERA: a community resource for metagenomics. *PLoS Biol*. 2007;5:e75. <https://doi.org/10.1371/journal.pbio.0050075>.
28. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One*. 2012;7: e47656. <https://doi.org/10.1371/journal.pone.0047656>.
29. Markowitz VM, Chen IM, 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(November):123–9.
30. Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P. SmashCommunity: A metagenomic annotation and analysis tool. *Bioinformatics*. 2010;26:2977–8.
31. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA - A practical iterative De Bruijn graph De Novo assembler. In: *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2010. p. 426–40.
32. Hyatt D, Chen G, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal : prokaryotic gene recognition and translation initiation site identification. 2010.
33. Moreno-Hagelsieb G. The power of operon rearrangements for predicting functional associations. *Comput Struct Biotechnol J*. 2015;13:402–6. <https://doi.org/10.1016/j.csbj.2015.06.002>.
34. 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.
35. 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.
36. Jacob E, Sasikumar R, Nair KNR. A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics*. 2005;21:1403–7.
37. Reese MG. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput Chem*. 2001; 26:51–6.
38. Weber T, Blin K, Duddela S, Krug D, Kim HU, Brucoleri R, et al. AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res*. 2015;43:W237–43.
39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(February):357–9.
40. 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.
41. Kim J, Kim MS, Koh AY, Xie Y, Zhan X. FMAP : Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics*. 2016;1–8. <https://doi.org/10.1186/s12859-016-1278-0>.
42. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*. 2012;40(Database issue):D109–14. <https://doi.org/10.1093/nar/gkr988>.
43. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004;32(suppl_1):D277–80. <https://doi.org/10.1093/nar/gkh063>.
44. 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*. 2015;31:926–32. <https://doi.org/10.1093/bioinformatics/btu739>.
45. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat eMethods* 2014;12:59. <https://doi.org/https://doi.org/10.1038/nmeth.3176>.
46. 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.
47. Chen X, Su Z, Xu Y, Jiang T. Computational Prediction of Operons in *Synechococcus* sp. WH8102. *Genome Inform*. 2004;15(2):211–22.
48. Bergman NH, Passalacqua KD, Hanna PC, Qin ZS. Operon prediction for sequenced bacterial genomes without experimental information. *Appl Environ Microbiol*. 2007;73:846–54.
49. Chuang L, Yang C, Tsai J, Yang C. Operon prediction using chaos embedded particle swarm optimization. *IEEE/ACM Trans Comput Biol Bioinform*. 2013;10(5):1299–309. <https://doi.org/10.1109/TCBB.2013.63>.
50. 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.
51. 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.
52. Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res*. 2010;38.
53. Vey G. Metagenomic guilt by association: an operonic perspective. *Plos One*. 2013;8(8):e71484.
54. 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*. 2012;490:55–60. <https://doi.org/10.1038/nature11450>.
55. 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.
56. Vey G, Charles TC. MetaProx: the database of metagenomic proximons. *Database*. 2014;2014:bau097–bau097. doi:<https://doi.org/10.1093/database/bau097>.

57. Vey G, Charles TC. An analysis of the validity and utility of the proximon proposition. 2012;215–20.
58. Detlev G, Vey A. The Proximon : Representation , Evaluation , and Applications of Metagenomic Functional Interactions by.
59. 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(Database issue):D394–7. <https://doi.org/10.1093/nar/gkj156>.
60. Mao F, Dam P, Chou J, Olman V, Xu Y. DOOR: A database for prokaryotic operons. *Nucleic Acids Res.* 2009;37(SUPPL. 1):459–63.
61. 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(December):288–98.
62. 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.
63. 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.
64. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S. Computational identification of operons in microbial genomes. *Genome Res.* 2002;12(8): 1221–30. <https://doi.org/10.1101/gr.200602>.
65. Rinninella E, Raoul P, Cintoni M, Franceschi F, Miggiano GAD, Gasbarrini A, Mele MC. What is the healthy gut microbiota composition? a changing ecosystem across age, environment, diet, and diseases. *Microorganisms.* 2019;7(1):14. <https://doi.org/10.3390/microorganisms7010014>.
66. Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. NeSSM: A Next-Generation Sequencing Simulator for Metagenomics. *PLoS One.* 2013;8.
67. Bratlie MS, Johansen J, Drabløs F. Relationship between operon preference and functional properties of persistent genes in bacterial genomes. *BMC Genomics.* 2010;28(11):71. <https://doi.org/10.1186/1471-2164-11-71>.
68. Price MN, Arkin AP, Alm EJ. The life-cycle of operons. *PLoS Genet.* 2006; 2(June):0859–73.
69. Nuñez PA, Romero H, Farber MD, EPC R. Natural selection for operons depends on genome size. *Genome Biol Evol.* 2013;5:2242–54.
70. Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. *Nucleic Acids Res.* 2001;29:1216–21.
71. 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 A Tox Hazard Subst Environ Eng.* 2015;50:1136–47.
72. 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.* 2013;5:3. <https://doi.org/10.1186/1758-5996-5-3>.
73. 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.
74. 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.* 2018;67(Supplement 1):1716. <https://doi.org/10.2337/db18-1716-P>.
75. 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.* 2015;64:1108–19. <https://doi.org/10.2337/db14-0646>.
76. 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 and Glycosaminoglycans as Clinical Biomarkers and Therapeutics - Part A.* Academic Press; 2019. p. 25–57. doi:<https://doi.org/https://doi.org/10.1016/b978-0-12-812001-1.00011>.
77. Wijnhoven TJ, van den Hoven MJ, Ding H, van Kuppevelt TH, van der Vlag J, Berden JH, Prinz RA, Lewis EJ, Schwartz M, Xu X. Heparanase induces a differential loss of heparan sulphate domains in overt diabetic nephropathy. *Diabetologia.* 2008;51(2):372–82. <https://doi.org/10.1007/s00125-007-0879-6>.
78. Yokoyama H, Sato K, Okudaira M, Morita C, Takahashi C, Suzuki D, Sakai H, Iwamoto Y. Serum and urinary concentrations of heparan sulfate in patients with diabetic nephropathy. *Kidney Int.* 1999;56(2):650–8. <https://doi.org/10.1046/j.1523-1755.1999.00591.x>.
79. Lauer ME, Hascall VC, Wang A. Heparan sulfate analysis from diabetic rat glomeruli. *J Biol Chem.* 2007;12;282(2):843–52. <https://doi.org/10.1074/jbc.M608823200>.
80. Bishop JR, Foley E, Lawrence R, Esko JD. Insulin-dependent diabetes mellitus in mice does not alter liver heparan sulfate. *J Biol Chem.* 2010;285(19): 14658–62. <https://doi.org/10.1074/jbc.M110.112391>.
81. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47(D1):D506-15. <https://doi.org/10.1093/nar/gky1049>.
82. 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.
83. Treadway JL, Mendys P, Hoover DJ. Glycogen phosphorylase inhibitors for treatment of type 2 diabetes mellitus. *Expert Opin Investig Drugs.* 2001; 10(3):439–54. <https://doi.org/10.1517/13543784.10.3.439>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

