

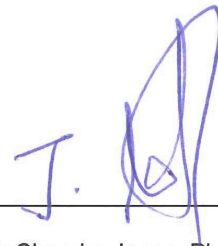
**OperomeDB: DATABASE OF CONDITION SPECIFIC TRANSCRIPTION IN  
PROKARYOTIC GENOMES AND GENOMIC INSIGHTS OF CONVERGENT  
TRANSCRIPTION IN BACTERIAL GENOMES**

Kashish Chetal

Submitted to the faculty of the Bioinformatics Graduate Program in partial fulfillment of  
the requirements for the degree Master of Science in Bioinformatics in the School of  
Informatics and Computing Indiana University October 2014

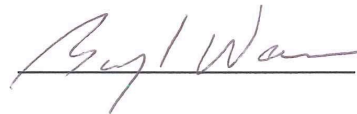
Accepted by the Graduate faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

**Master's Thesis Committee**



---

Sarath Chandra Janga, Ph.D.



---

Barry L Wanner, Ph.D.



---

Yunlong Liu, Ph.D.

**Copyright page**

© 2014

Kashish Chetal

ALL RIGHTS RESERVED

To my parents, Mr. Pardeep Chetal and Mrs. Usha Chetal and my family. They raised me, supported me, taught me and loved me. This thesis is dedicated to them.

## **Acknowledgement**

This thesis was made possible due to the guidance and encouragement from many people. So it gives me a great pleasure to thank these people and acknowledge their contribution. I owe a sincere appreciation to my thesis advisor Dr Sarath Chandra Janga for his motivation and immense knowledge. Without his thoughtful guidance, energy and constructive criticism this thesis would have never been possible. Besides my lab members, I would also like to thank the other members of my committee Dr Barry L Wanner and Dr Yunlong Liu for supporting my work, reading my thesis and providing helpful suggestions.

I wish to express sincere appreciation to School of Informatics at Indiana University Purdue University Indianapolis for providing me an opportunity to pursue a bright carrier in bioinformatics.

I thank my family for the constant support, love and blessing. Their teachings have made me the person I am today.

Kashish Chetal

**OperomeDB: DATABASE OF CONDITION SPECIFIC TRANSCRIPTION IN  
PROKARYOTIC GENOMES AND GENOMIC INSIGHTS OF CONVERGENT  
TRANSCRIPTION IN BACTERIAL GENOMES**

Abstract

My thesis comprises of two individual projects: 1) we have developed a database for operon prediction using high-throughput sequencing datasets for bacterial genomes. 2) Genomics and mechanistic insights of convergent transcription in bacterial genomes.

In the first project we developed a database for the prediction of operons for bacterial genomes using RNA-seq datasets, we predicted operons for bacterial genomes. RNA-seq datasets with different condition for each bacterial genome were taken into account and predicted operons using Rockhopper. We took RNA-seq datasets from NCBI with distinct experimental conditions for each bacterial genome into account and analyzed using tool for operon prediction. Currently our database contains 9 bacterial organisms for which we predicted operons. User interface is simple and easy to use, in terms of visualization, downloading and querying of data. In our database user can browse through reference genome, genes present in that genome and operons predicted from different RNA-seq datasets.

Further in the second project, we studied the genomic and mechanistic insights of convergent transcription in bacterial genomes. We know that convergent gene pairs with overlapping head-to-head configuration are widely spread across both eukaryotic and prokaryotic genomes. They are believed to contribute to the regulation of genes at both transcriptional and post-transcriptional levels, although factors contributing to their

abundance across genomes and mechanistic basis for their prevalence are poorly understood. In this study, we explore the role of various factors contributing to convergent overlapping transcription in bacterial genomes. Our analysis shows that the proportion of convergent overlapping gene pairs (COGPs) in a genome is affected due to endospore formation, bacterial habitat, oxygen requirement, GC content and the temperature range. In particular, we show that bacterial genomes thriving in specialized habitats, such as thermophiles, exhibit a high proportion of COGPs. Our results also conclude that the density distribution of COGPs across the genomes is high for shorter overlaps with increased conservation of distances for decreasing overlaps. Our study further reveals that COGPs frequently contain stop codon overlaps with the middle base position exhibiting mismatches between complementary strands. Further, for the functional analysis using cluster of orthologous groups (COGs) annotations suggested that cell motility, cell metabolism, storage and cell signaling are enriched among COGPs, suggesting their role in processes beyond regulation. Our analysis provides genomic insights into this unappreciated regulatory phenomenon, allowing a refined understanding of their contribution to bacterial phenotypes.

# Contents

Chapter 1 Introduction.....	7
1.1 What is Gene Regulation? .....	7
1.2 Operons in bacterial genomes: analysis and prediction using RNA-seq datasets ..	9
1.3 Genomic and mechanistic Insights of convergent transcription in bacterial genomes.....	10
Chapter 2 OperomeDB: a database of condition-specific transcription units in prokaryotic genomes .....	12
2.1 Introduction.....	12
2.2 Datasets .....	17
We collected RNA-seq datasets for various bacterial species under a number of different conditions from the Sequence Read Archive (SRA) of NCBI (Kodama, et al.) as described below.....	17
Chapter 3 Genomic and mechanistic Insights of convergent transcription in bacterial genomes .....	31
3.1 Introduction.....	31
3.2 Material and Methods.....	33
3.3 Results and discussion.....	35
Chapter 4 Conclusion .....	53
Chapter 5 Future Work .....	55
Chapter 6 References.....	56



## List of Figures

Figure 1. Web interface for operomeDB showing a screenshot of a selected bacterial genome to facilitate the browsing and download of predicted operons. ...	26
Figure 2. Screenshot showing the selection of an operon in the operons track. Highlighted is the ecpBCDE operon in <i>Escherichia coli</i> K12 genome encoding for the membrane and fimbria formation proteins. ....	27
Figure 3. Snapshot of the jBrowse visualization showing the ensemble of all the operons predicted for a bacterial organism. ....	28
Figure 4. Presence and absence of operons for different experimental conditions in two different bacterial genomes. ....	29
Figure 5. Operonic view showing a newly identified operon (yeaP-yoaK-yoaJ) in the genome of <i>Escherichia coli</i> K12. In operomeDB, newly identified operons compared to other databases such as DOOR are marked as 'NA' and user can further click on these to get the relevant information. ....	30
Figure 6. Flowchart for methodology and processing of bacterial genome data. ...	36
Figure 7. Distribution of proportion of COGPs across genome size for bacterial genome groups. ....	38
Figure 8. Comparison of proportion of COGPs to different lifestyle for bacterial genomes. ....	39
Figure 9. Density distribution plots for various lifestyles on the basis of their proportion. ....	41
Figure 10. Density distribution of bacterial genomes for COGPs. ....	44
Figure 11. Conservation of COGPs across shorter overlaps for all genomes. ....	46

Figure 12. Functional enrichment of convergent gene pairs..... 47

Figure 13. Sequence logo representation for gene sequence..... 51

## List of Tables

Table 1. Detailed p-value calculated using paired Wilcoxon test for the proportion of COGPs across bacterial genomes for different lifestyles .....	41
Table 2. P-values for individual and combined factors for different lifestyles calculated using anova. Proportion of COGPs was taken as a response variable to other lifestyles, and p-values were calculated for each.....	43
Table 3. Percentage and mismatch of stop codons for COGPs when mapped to gene sequence (A) determines the COGPs for inter-genic distance less than $<-4$ (B) It determines the COGPs for inter-genic distance $-4$ .....	49
Table 4. Percentage of each stop codon across COGPs (A) for intergenic distance $-4$ , (B) for intergenic distance less than $<- 4$ .....	50
Table 5. This table contains the significant p-values for individual factors that are significant for a particular lifestyle. Adjusted R-squared value and p-value were also calculated using multiple regression, which shows all the factors significant for bacteria with COGPs. ....	52

## Chapter 1 Introduction

### 1.1 What is Gene Regulation?

Bacteria are the simplest form of free-living life known to man. They are single-celled and vulnerable to adverse and dynamic environmental forces and yet, they have colonized diverse niches. Bacteria can thrive in any different environment and habitats. Bacteria are capable of living on a single host or large host, different habitats, different environmental conditions and to the extent of animal host. The necessity of living in diverse habitats forces bacteria to form molecular tools to survive under these conditions (Seshasayee, et al.).

Bacteria do not make all the protein as which they are capable of making. Instead of that they adapt to the particular environment in which they are living and make only those gene products, which are essential for them to survive in particular condition. There are some of the gene products required by bacteria to survive in any condition and those are called housekeeping genes (Seshasayee, et al.). This includes the genes that encode different protein such as DNA polymerase, RNA polymerase and DNA gyrase. For example, if the tryptophan is present in abundance in environment then bacteria will not produce the enzymes, which help in production of tryptophan, whereas if the former is not present in the environment bacteria will produce the enzyme, which produces tryptophan. So bacteria generally control the expression of gene by regulating the process of mRNA transcription (Winkler and Breaker). Here we will discuss about expression regulatory aspects in bacteria, as how it regulates the expression of its genes so that which genes will be expressed and how they will control the expression or cell growth conditions. There are many studies carried out to understand that how bacteria regulate its expression in response to different extracellular and intracellular conditions (Browning and Busby). There are hundreds

of different transcription factor found due to large scale sequencing of bacterial genomes, so regulation of transcription is a key factor for bacteria to regulate its gene expression (Scheffers and Errington).

Regulation of gene can occur at different places and time points to produce an active gene product. Gene can be regulated using transcriptional machinery or translational machinery, when the gene is transcribed and how much it is transcribed would tell about its expression and this is called as transcriptional regulation. Gene products are also regulated when they are completely synthesized at post-transcriptional and post-translational level. Post transcription includes those that control transcription elongation, transcription termination, translation initiation, and translational termination mechanisms (Babitzke) (Stulke).

When two convergent promoters locate on a DNA it's called convergent transcription (Crampton, et al.). Convergent transcription is a simultaneous induction of the sense and antisense transcription through two different opposing promoters (Lin, et al.). Sense and antisense transcripts frequently occurs in prokaryotic and eukaryotic organisms, so convergent transcription provides a meaningful role in the process of gene expression and in the process of functionality. Many studies have been done to document the role of convergent transcription to provide the evidence about the functionality and gene regulation occurring in different organisms. Convergent gene transcription also allows understanding of the biology and the process of transcriptional gene silencing by involving RNA interference mechanism (Gullerova and Proudfoot). Bacterial genes are organized into a cluster of genes called operons, which are co-regulated, and these all are controlled by the same promoter. In various studies it has been discussed that in various bacterial species

the structure of operons change with the environment (Guell, et al.). So every step, which requires preparing an active gene product, leads to gene regulation.

## 1.2 Operons in bacterial genomes: analysis and prediction using RNA-seq datasets

We present OperomeDB (<http://sysbio.informatics.iupui.edu/operomeDB/>), which provides an ensemble of all the predicted operons for bacterial genomes using available RNA-sequencing datasets across a wide-range of experimental conditions(Guell, et al.). Although several studies have recently confirmed that prokaryotic operon structure is dynamic with significant alterations across environmental and experimental conditions, there are no comprehensive databases for studying such variations across prokaryotic transcriptomes. To address this gap, we exploited the growing number of publicly available RNA-sequencing datasets from NCBI-SRA for various experimental conditions across diverse bacterial genomes, to provide a one stop portal for understanding the genome organization in the context of transcriptional regulation in a condition-specific manner. Currently our database contains nine bacterial organisms and 168 transcriptomes for which we predicted operons. User interface is simple and easy to use, in terms of visualization, downloading and querying of data. Users can browse through the reference genome, genes and operons predicted in the genome based on RNA-seq datasets as well as those identified in specific conditions. In addition, because of its ability to load custom datasets, users can also compare their datasets with publicly available transcriptomic data of an organism.

OperomeDB as a database, should not only aid experimental groups working on transcriptome analysis of specific organisms but also enable studies related to computational and comparative operomics.

Our database can be assessed at <http://sysbio.informatics.iupui.edu/operomeDB/>,

### 1.3 Genomic and mechanistic Insights of convergent transcription in bacterial genomes

Convergent gene pairs with overlapping head-to-head configuration are widely spread across both eukaryotic and prokaryotic genomes. They are believed to contribute to the regulation of genes at both transcriptional and post-transcriptional levels, although factors contributing to their abundance across genomes and mechanistic basis for their prevalence are poorly understood. In this study, we explore the role of various factors contributing to convergent overlapping transcription in bacterial genomes. Our analysis shows that the proportion of convergent overlapping gene pairs (COGPs) in a genome is affected due to endospore formation, bacterial habitat, oxygen requirement, GC content and the temperature range. In particular, we show that bacterial genomes thriving in specialized habitats, such as thermophiles, exhibit a high proportion of COGPs. Our results also show that the density distribution of COGPs across the genomes is high for shorter overlaps with increased conservation of distances for decreasing overlaps. Our study further reveals that COGPs frequently contain stop codon overlaps with the middle base position exhibiting mismatches between complementary strands. Functional analysis using cluster of orthologous groups (COGs) annotations suggested that cell motility, cell metabolism, storage and cell

signaling are enriched among COGPs, suggesting their role in processes beyond regulation. In conclusion our study provides genomic insights into this unappreciated regulatory phenomenon, allowing a refined understanding of their contribution to bacterial phenotypes.

For this project, we used 2,168 bacterial genomes to predict their functionality, lifestyle, distribution and conservation for convergent overlapping gene pairs across transcription. For the study, we first paired the genes as FF or RR, FR and RF; from this we selected the gene pairs with FR strands, which are responsible for convergent transcription. Gaussian density distribution was used to predict the highest density distribution point and the inter-genic distance for each genome having COGPs. This led us to find the conservation pattern across all genomes, where most of the convergent transcription occurs. To predict functionality, we used COGs and mapped them to the FR strand to anticipate their functionality across convergent gene pairs.



## Chapter 2 OperomeDB: a database of condition-specific transcription units in prokaryotic genomes

### 2.1 Introduction

As the gap between the rate at which sequencing of complete genomes and the experimental characterization of transcriptional regulation in them increases, automated computational methods for unravelling the regulatory code are increasingly being sought after. Although accurate tools for identifying the genes encoded in a genome have been developed, our understanding on how the genes are expressed and regulated depends on our knowledge of how they are organized into operons - sets of genes that are co-transcribed to produce a single messenger RNA (Jacob, et al., 1960; Jacob, et al., 2005). Operons are the essential units of transcription in prokaryotic organisms, and as a result, identifying these structures is a main step in understanding transcriptional regulation. Knowing operon structure in a genome not only facilitates to identify sets of genes, which are co-regulated but also aids in other computational analyses, such as prediction of cis-regulatory elements which often depend on accurate detection of operons. In addition, since operons often consist of genes that are related functionally and required by the cell for a numerous biological process, they are often good predictors of biological modules (Dandekar, et al., 1998; Janga, et al., 2005; Overbeek, et al., 1999). Therefore, deep understanding of operons, will improve our knowledge of higher-order genomic associations and structures thereby expanding our understanding of various cellular networks composed of regulatory, structural and functional pathways (Janga, et al., 2005; Lathe, et al.). Operons also provide insights into the cellular functions and also help in determining different experimental designs. In various recent high-throughput RNA-sequencing studies across a number of prokaryotic organisms, it has been convincingly shown that the structure of operons' changes

with the environmental conditions (Guell, et al.; Sorek and Cossart). Thus, suggesting a need to the discovery and a better understanding of the transcriptional units originating from operons (predicted or otherwise) across experimental conditions in bacterial genomes. For all these reasons, the characterization of condition-specific transcription unit structure on a genomic scale is an important starting point for microbial functional genomics.

Several operon databases are currently available and provide information with varying levels of reliability and emphasis (Chivian, et al.; Mao, et al.; Pertea, et al.; Salgado, et al.; Taboada, et al.). However, it is important to note that traditionally, definitions of operons and transcription units are synonymously used for computational predictions, mainly because each operon was believed to encode for a single transcription unit (single polycistronic unit). However, emerging evidence from several RNA-sequencing studies support a more complex model, with several operons in a genome encoding for multiple transcription units depending on the condition (Guell, et al.; Sorek and Cossart). Databases such as RegulonDB (Salgado, et al.), which are based on manual curation of experimentally reported polycistronic transcripts identified in at least one experimental condition in the literature in *Escherichia coli* K12, define an operon as the ensemble of all the transcription units in a given genome loci which results in the longest stretch of co-directional transcript. In such frameworks, each transcription unit is governed by a promoter and terminator identified in atleast one condition. In contrast, working definition for computational prediction of operons across most studies simply assumes the longest possible polycistronic transcript in a genomic locus as an operon. These differences in the working definition indicate that the current prediction pipelines and databases for operon prediction are from perfect in

predicting condition-specific transcription units/operons in bacterial genomes. In OperonDB, Perteau et al. employed a method to find and analyze gene pairs that are located on the same strand of DNA in two or more bacterial genomes (Perteau, et al.). The computational algorithm used in this database locates operons structure in microbial genomes using a method published earlier by the authors (Ermolaeva, et al.). OperonDB currently contains 1059 genomes with prediction sensitivity of 30%-50% in *Escherichia coli* (Perteau, et al.). DOOR (database for prokaryotic operons) is another database, which contains predicted operons for 675 sequenced prokaryotic genomes. It provides similarity scores between operons by which user can search for related operons in different organisms (Mao, et al.). ProOpDB (prokaryotic operon database), predicts operons in more than 1200 prokaryotic genomes using a neural network based approach. It provides several options for retrieving operon information. In ProOpDB, users can also visualize operons in their genomic context and their nucleotide or amino acid sequences (Taboada, et al.).

MicrobesOnline is another operon database, which facilitates the phylogenetic analysis of genes from microbial genomes (Chivian, et al.). In principle, this database has two functionalities 1) user can build a phylogenetic tree for every gene family as well as a species tree in a tree-based browser to assist in gene annotation and in reconstructing their history of evolution, 2) using its tool one can analyze microarray data to find genes which exhibit similar expression profiles in an organism which can subsequently be used for identifying regulatory motifs and seeing if they are conserved. User can also compare the organization of a protein domain with genes of interest in a browser (Chivian, et al.). Finally, as mentioned above, RegulonDB is a database (Salgado, et al.), which is curated and designed for *Escherichia coli K12* to facilitate the prediction of its transcriptional regulatory network and operon

organization across growth conditions. It also provides extensive information about the evolutionary conservation of a number of regulatory elements in *Escherichia coli* genome. . The method used in RegulonDB has a certainty of 88% identification of pairs of genes, which are adjacent in operon, and it also describe 75% of the known transcription units which are used to predict the transcriptional organization of *Escherichia coli* genome (Salgado, et al.). However, there is not a single database present, which uses data from RNA-sequencing experiments to predict the transcription unit organization in a broad range of bacterial genomes in a condition-specific manner. In this study, we present operomeDB to address this gap – a database dedicated to the identification and visualization of transcriptional units from publicly available RNA-seq data in microbial genomes.

High-throughput sequencing platforms like illumina, ABI and Roche are used to quantify the expression levels of RNA in a condition-specific manner in bacterial genomes –frequently referred to as an RNA-seq experiment. Such high-throughput technologies have several advantages compared to traditionally used microarray platforms like a low background signal, large dynamic range of expression level, and possibility of detecting novel transcripts. There are different tools for detection, management and analysis of the eukaryotic RNA-seq data, however relatively very few tools are available for the analysis and processing of RNA-seq data in prokaryotes. Rockhopper is an open source computational algorithm implemented for the analysis of bacterial RNA-seq data (McClure, et al.). It supports different stages of RNA-seq analysis and datasets from different sequencing platforms. The algorithm performs several functions such as aligning the sequence reads to a genome, constructing transcriptome maps, calculating the abundance of transcripts, differential gene expression and predicting transcription unit structure. It also has the

ability to detect novel small RNAs, operons and transcription start sites with a high accuracy in a transcriptome specific manner (McClure, et al.).

Although there are many tools and software's available for the visualization and exploration of next-generation sequencing datasets for eukaryotic organisms, there is a lack of proper genome browsers to visualize prokaryotic organisms and transcriptomes in particular. The size of data generated by RNA sequencing methods is usually large and makes data visualization a challenging task. IGV (Integrative genomic viewer) is a visualization tool that can visualize large data sets very smoothly with the main aim of helping the researchers to visualize and explore the results (Thorvaldsdottir, et al.). The UCSC and Ensembl genome browsers are online tools that have been used to display different biological datasets, including genomic variants, expressed sequence tags and functional genomic data with manually curated annotations (Flicek, et al.; Goldman, et al.). In this study, we used jBrowse to develop visualization of predicted transcription units for each RNA-seq dataset analyzed across genomes. jBrowse is an open source, portable, JavaScript based genome browser particularly suitable for prokaryotic genomes. The browser provides easy navigation of genome annotations on the web and has good track selection, zooming, panning and navigation features (Skinner, et al.).

We believe that biological community could benefit from having a new operon prediction database, which uses RNA-seq datasets to predict transcription units in a condition/transcriptome-specific manner. In our presented database (operomeDB) for bacterial genomes, we used an innovative approach to query operons. We predict operons for nine bacterial genomes for which at least few RNA-seq datasets are available in the public domain from the Sequence Read Archive (SRA) of NCBI

(Kodama, et al.). We used Rockhopper (McClure, et al.) for the computational analysis of data. Using RNA-seq data for different bacterial genomes, the developed database, which to our knowledge is the largest of its kind to date, should facilitate researchers to navigate through operons predicted under different experimental conditions.

## 2.2 Datasets

We collected RNA-seq datasets for various bacterial species under a number of different conditions from the Sequence Read Archive (SRA) of NCBI (Kodama, et al.) as described below.

### **Escherichia coli K-12 Mg1655**

*Escherichia coli* is generally found in the colon and large intestine of the warm-blooded organisms. It belongs to a family of k-12 and B strain that is used in molecular biology for different experiments and also considered as a model organism. K-12 is the strain first confined from a sample of stool of the patient suffering from diphtheria. Different strains have been emerged in years due to various treatment agents (Stothard, et al.). Expression profiling of wild type and SgrR mutant *E. coli* under aMG and 2-DG-induced strain were performed by Wadler et. al (Wadler and Vanderpool). RNA-sequencing data available for illumina platform for this strain under 54 different conditions was analyzed using Rockhopper (McClure, et al.).

### **Eggerthella Lenta DSM2243**

*Eggerthella lenta* is an anaerobic, non-motile, non-sporulating pathogenic gram-positive bacteria confined from rectal tumor. It is mostly found in blood and human intestine and can cause severe infections. Temperature favorable for growth of these bacteria is 37 degree Celsius (Stothard, et al.). Expression profiling study carried out for the generation of datasets is based on RNA-Seq analysis of *Eggerthella lenta* cultured with or without digoxin. This dataset comprised of 21 different transcriptomes in *Eggerthella lenta* DSM2243 strain from Haiser et. al (Haiser, et al.).

### **Campylobacter Jejuni RM1221**

Campylobacter species are the prominent cause of gastroenteritis in countries on the path of development. An infection occurring due to *C. jejuni* is the most frequent preliminary cause for a neuromuscular paralysis, which is also known as Guillain-Barre syndrome. Healthy cattle and birds can carry *C. jejuni* (Stothard, et al., 2005). For this study, data from Dugar et al. (Dugar, et al.) did the comparative dRNA-seq analysis of multiple campylobacter jejuni strains revealed a conserved and specific to strain transcription pattern was used. For 16 different conditions RNA-seq data for *Campylobacter jejuni* RM1221 was obtained from this study (Dugar, et al.).

### **Clostridium Beijerincki NCIMB 8052**

*C. beijerinckii* NCIMB 8052 is anaerobic, motile, rod-shaped bacteria. The anatomy of the cell changes with the progression of growth cycle of the organism. *C. beijerinckii* species are present everywhere in nature and routinely segregated from soil samples (Stothard, et al., 2005). Wang et al. carried out single-nucleotide resolution analysis of the transcriptomic structure of *Clostridium beijerincki* NCIMB 8052 using RNA-seq technology (Wang, et al.). This comprised of expression quantification dataset for 6 different conditions in this organism (Wang, et al.).

### **Clostridium difficile 630 RNA-seq experiments**

*C. difficile* is commonly found in water, air, human and animal feces. Its genome reveals that the pathogen thrives in the gastrointestinal tract and some of its strains are more fatal than others. With the help of *C. difficile* genome we can understand the antimicrobial resistance and various treatment options available. After the sequencing of the whole genome, researchers found that from the whole genome, 11% of it consists of genetic elements such as conjugative transposons. These genetic elements contribute *clostridium* with the genes subjected for their antimicrobial resistance, interaction to host and surface structure production (Sebaihia, et al., 2006). We used data from Fimlaid et al., where the authors conducted a global analysis of genes induced during sporulation of *Clostridium difficile* using Illumina HiSeq 1000 for 18 different conditions (Fimlaid, et al.).

### **Mycobacterium tuberculosis H37rv**

*Mycobacterium* is a causative agent of tuberculosis and has a waxy coating on its surface. Primary *mycobacterium* affects respiratory system and lungs. H37rv strain of tuberculosis has 4 million base pairs with 3959 genes. The genome contains 250 genes that are involved in metabolism of fatty acids. Datasets for this genome are collected from experiments in which authors performed the high-resolution transcriptome and genome wide dynamics of RNA polymerase and NusA (Uplekar, et al.). A total of 10 different transcriptomes were collected from this study for *Mycobacterium tuberculosis* (Uplekar, et al.).

### **Salmonella enterica subsp. enterica serovar typhimurium str. 14028S**



*Salmonella enterica* serovar is a subspecies of *S. enterica*, these are in the shape of rod, flagellated, aerobic and gram-negative. *Salmonella* serovar can have many strains, which allows for accelerated increase in the total number of antigenically variable bacteria. In a study by Stringer et. al (Stringer, et al.), authors used RNA-seq to conclude the effects of AraC and arabinose on RNA levels genome-wide in *S. enterica*. Wild type or delta AraC mutant cells were developed in the presence and absence of 0.2% L-arabinose. The data for *Salmonella enterica* was collected for 8 different conditions (Stringer, et al.)

### ***Sinorhizobium meliloti* 2011**

*Sinorhizobium meliloti* is a nitrogen-fixing bacterium. Nitrogen fixation by *S. meliloti* is hampered by the plastic modifier bisphenol A. Dataset used in our database corresponded to a recent study where the authors performed RNA-sequencing of 18 samples corresponding to this bacteria in 3 different conditions (Sallet, et al.). For each condition, both short and long RNA fractions were analyzed, and three replicates per condition and per RNA fraction were performed. In this study next generation annotation of prokaryotic genomes with EuGene-P was performed - applied to *Sinorhizobium meliloti* 2011 genome (Sallet, et al.).

### ***Synechococcus elongatus* PCC 7942**

*Synechococcus elongatus* are found in aquatic environments. They are called photosynthetic bacteria, as they are responsible for its production. *Synechococcus* consists of one circular chromosome and two plasmids. This particular strain contains a circular chromosome 2,700,000 bp long with GC content of 55 %. For the generation of 17 datasets, three strains (7942, SE01 and SE02) were analyzed by Ruffing at two time points (100 h and 240 h) with three biological replicates (Ruffing).

### 2.3 Prediction of operons using Rockhopper

To predict transcription units (operons) in a condition/transcriptome-specific manner, we used Rockhopper, a computational algorithm which supports different stages of RNA-seq analysis for datasets originating from diverse sequencing platforms (McClure, et al.). Rockhopper takes sequenced RNA reads as input in a number of formats including FASTQ, QSEQ, FASTA, SAM and BAM files (McClure, et al.). It allows the processing of next-generation RNA-seq data by permitting the user to specify different parameters to align sequence reads to a genome, such as number of mismatches allowed, orientation of mate-pair reads and minimum seed length. For transcriptomic analysis in Rockhopper, some parameters specified include whether the dataset is strand specific, test for differential expression, prediction of operons and minimum expression of UTRs and detection of ncRNAs. However, the authors recommend the use of default settings most of the time for best operon prediction performance and hence in this study we used the default parameters where possible (McClure, et al.). Indeed, operon prediction by Rockhopper has been shown by the original authors to perform at ~90% accuracy when benchmarked against RegulonDB (Salgado, et al.) and DOOR (Mao, et al.) databases. Each run of Rockhopper on a single RNA-seq dataset corresponding to a condition, provides different files as output, such as summary file - which contains a summary analysis of successfully aligned reads to genomic regions, transcript file - which includes newly predicted transcripts, transcription start and stop sites with expression levels. Finally, it provides operons file containing predicted operons in the condition. We ran Rockhopper in a batch mode to process and predict operons in each condition for each genomic dataset discussed above by selecting the appropriate reference

sequence. We also ran operon prediction on the complete transcriptomic dataset for each genome to obtain a consensus set of operon predictions which was used to show as a reference operome (operon track) of the organism in jBrowse. In order to index the operons in our database, we numbered them by matching with the IDs of the predicted operons from DOOR database in order to easily know the novel operons. If our predicted operon shared at least one gene we gave the same operon ID as DOOR database and for operons, which were not present in DOOR database, we marked them as 'NA'.

## 2.4 Visualization using jBrowse

In our database we incorporated jBrowse, which supports different file formats and in our specific implementation, we use FASTA files to display the reference sequence and BED, GFF or BAM format files for displaying the list of genes and other discrete features such as operons (Westesson, et al.). User can select the particular operon and selecting that particular operon can display the length of the operon, genes constituting the specific operon as well as sequence for that particular operon (Fig.3). From jBrowse panel, user can also select any number of experimental conditions for which operon predictions using RNA-seq data are available, and it will display the operons for selected location. Users have the choice to display any number of tracks and visually compare them for downstream analysis. For instance, Figure 4 shows examples of predicted presence and absence of operons for different experimental conditions in *Escherichia coli* K12 MG1655 and *Campylobacter jejuni* RM1221. It was found that certain operons in microbial genomes studied here; were missing for a few experimental conditions. In our database we represent this variability of tracks with respect to the reference genome.

For example, in *Escherichia coli* K12 MG1655, '3025' operon encoding for the genes *speD* (S-adenosylmethionine decarboxylase) and *speE* (spermidine synthase) was found to be missing in the experimental condition SRX254733 (Figure 4A). Such observations could be contributed due to the specific experimental condition which the experimentalists interested in the operon can explore further on a case-by-case basis. Another example shown in Figure 4B from the *Campylobacter jejuni* RM1221 transcriptome also exhibits variability in operon organization. In this organism we found that '61693' operon (encoding for the poorly annotated ORFs CJE0054 and CJE0055) is missing in SRX155620. In our database multi-gene operons are predicted based on the co-transcription occurring in genes. Hence, the operons with a lack of occurrence of co-transcription would be identified as missing operons suggesting either a functional relevance of their absence or in few cases for very low abundant genes due to the lack of sequencing depth, in certain experimental conditions under study. We anticipate that with increase in the depth and number of conditions for which RNA-seq datasets will become available, it will become easy to tease functionally important condition-specific transcription units via operomeDB.

Our system also allows a user to submit their own sequence in specific file format and database will display its contents as an additional track. Using option in jBrowse, user can easily upload their data files to jBrowse or paste URLs, where data is present to display its contents. Various file formats such as GFF3, BigWig, BAM index, BAM and VCF are supported. User can also visualize and compare different tracks and hence analyze if there are similarities/dis-similarities between tracks. This feature will enable the comparison of new RNA-seq data for a given organism with already available public data for various experimental conditions available in operomeDB. Additionally, custom tracks will also enable comparison of operon

tracks of different closely related organisms to study the variations in transcript architecture across the length of the genome.

In comparison to earlier resources of bacterial operons, our database offers high quality single nucleotide resolution bacterial operon predictions based on high-throughput data sets.

## 2.5 Using the database: an example

Below we provide an example illustrating the functionality of operomeDB. The presented example (Figure 5) is from *Escherichia coli* K12 MG1655 genome for a newly identified three gene operon *yeaP-yoaK-yoaJ* which has not been annotated in other databases such as DOOR (Mao, et al.) highlighting novel predicted operons that can be identified and visualized using our database.

1. A user can go to the main page of our Graphical User Interface (GUI), click on 'Select Organism' and then it will provide the list of the entire bacterial organisms present in the database. For instance, selecting the query genome as *Escherichia coli* K12 MG1655 will display the page showing the operon predictions in various formats for *E. coli*.
2. On the query result page, it will display the information regarding *E. coli* and other possible options available. By clicking on the link 'View in jBrowse' will enable the user to navigate the data via genome browser through different tracks.
3. In genome browser on the left panel user can select any number of available tracks and selected tracks will be displayed in the browser window. The user can now go through each track and query different operons predicted in our database.
4. Using the download button user can download the fasta sequence file for a particular operon.
5. By selecting 'file' option in upper panel, users can also upload/add their own sequence or dataset for visualization or comparison in the genome browser.
6. User can also look for predicted operons in each bacterial organism marked as 'NA'. These are the operons that are newly predicted in our study compared to the DOOR operon database (Mao, et al.) (Figure 5).

OperomeDB's functionality may not be optimum in mozilla firefox where there are known issues reported for jBrowse. We anticipate resolving these issues with the

help of the developers of jBrowse in the next release of operomeDB so that the database is accessible in all platforms and browser

## 2.5 Implementation and Interface

We developed an interface using HTML, CSS, JavaScript and also incorporated jBrowse, a genome viewer to display different tracks for building operomeDB (<http://sysbio.informatics.iupui.edu/operomeDB/>) presented here. User interface of our database is shown in Figure 1 which allows the selection of an organism using a drop-down list. User can select an organism and selected bacterial genome information is displayed. There are multiple view options available for users, like viewing as a table of predicted operons, viewing operons using jBrowse or to download the predicted operons as a table (Figure 1). Clicking on the tab with the option of 'view in jBrowse', will display data in jBrowse and user can view reference sequence of the genome, gene list in the particular bacterial genome and a list of operons predicted. User can also select to show the operon data for different conditions from which RNA-seq datasets have been taken, with reference to their SRA IDs. Also, using SRA ID, user can search for a specific condition of each bacterial genome in the NCBI SRA database (<http://www.ncbi.nlm.nih.gov/sra>) (Kodama, et al.).

For a selected operon or gene, jBrowse will provide detailed information such as genomic position of that particular operon, its length in base pairs (bp) and its primary attributes such as IDs, associated gene names, source and sequence region in FASTA format (Figure 2). For a selected gene in the gene track, additional attributes such as Dbxref (reference id) and Gbkey (CDS, Gene) are also displayed. Our database can generate a fasta file containing user-specified operons and

associated information and can be downloaded to the user's local computer for further analysis.

Operon Prediction for Bacterial Genomes DB

Select organism  
Escherichia Coli: K12 MG1655

Home  
View Tutorial  
Contact Us

## Welcome

Bacterial Operon Database consists of various bacterial genomes. In this database we have collected datasets for different bacterial genomes from their RNA-seq datasets. The operon predictions for these datasets are different from the normal operon predictions for various bacterial organisms. Many new operons have also been predicted from the RNA-seq datasets.

### Escherichia Coli K12 Mg1655

Escherichia coli is commonly found in the lower intestine of the warm-blooded organisms. It is the descendant of k-12 and B strain that is used in molecular biology for experimentation and tool as a model organism. K-12 is the strain isolated from a stool sample of the patient suffering from diphtheria. As now a days different strain have been emerging as treating it with various agents. Expression profiling of wild-type and SgrR mutant E.coli under aMG and 2-DG-induced strain has done. The RNA-seq analysis was one using Illumina for strain grown in different conditions.

<http://www.ncbi.nlm.nih.gov/sra?term=SRP019956>

View in jBrowse Show Operon Table Download Operon Table

**Figure 1. Web interface for operomeDB showing a screenshot of a selected bacterial genome to facilitate the browsing and download of predicted operons.**

The left panel of the webpage allows user to select an organism of interest. Once the user selects a bacterial organism the interface will provide information about the organism, experimental conditions under which RNA-seq datasets are available, SRA link for experimental conditions and options to visualize in jBrowse, show operon table and download the complete set of operon predictions across all the conditions as a table.

Select organism  
Escherichia Coli K12 MG1655

Home  
View Tutorial  
Contact Us

## Escherichia Coli K12 Mg1655

Go Back

Available Tracks

filter by text

- Gene
- Operon
- SRX254733
- SRX254734
- SRX254735
- SRX254736
- SRX254737
- SRX254738
- SRX254739
- SRX254740
- SRX254741
- SRX254742
- SRX254743
- SRX254744
- SRX254745
- SRX254746
- SRX254747
- SRX254748
- SRX254749
- SRX254750
- SRX254751
- SRX254752
- SRX254753
- SRX254754
- SRX254755
- SRX254756
- SRX254757
- SRX254758
- SRX254759
- SRX254760
- SRX254761
- SRX254762
- SRX254763

3060

**Primary Data**

<b>Name</b>	3060
<b>Type</b>	3060
<b>Position</b>	gi 556503834 ref NC_000913.3 :303719..309250 (- strand)
<b>Length</b>	5,532 bp


**Attributes**

<b>Id</b>	3060
<b>Names</b>	ecpE ecpD ecpC ecpB
<b>Seq_id</b>	NC_000913.3
<b>Source</b>	RefSeq

**Region sequence**

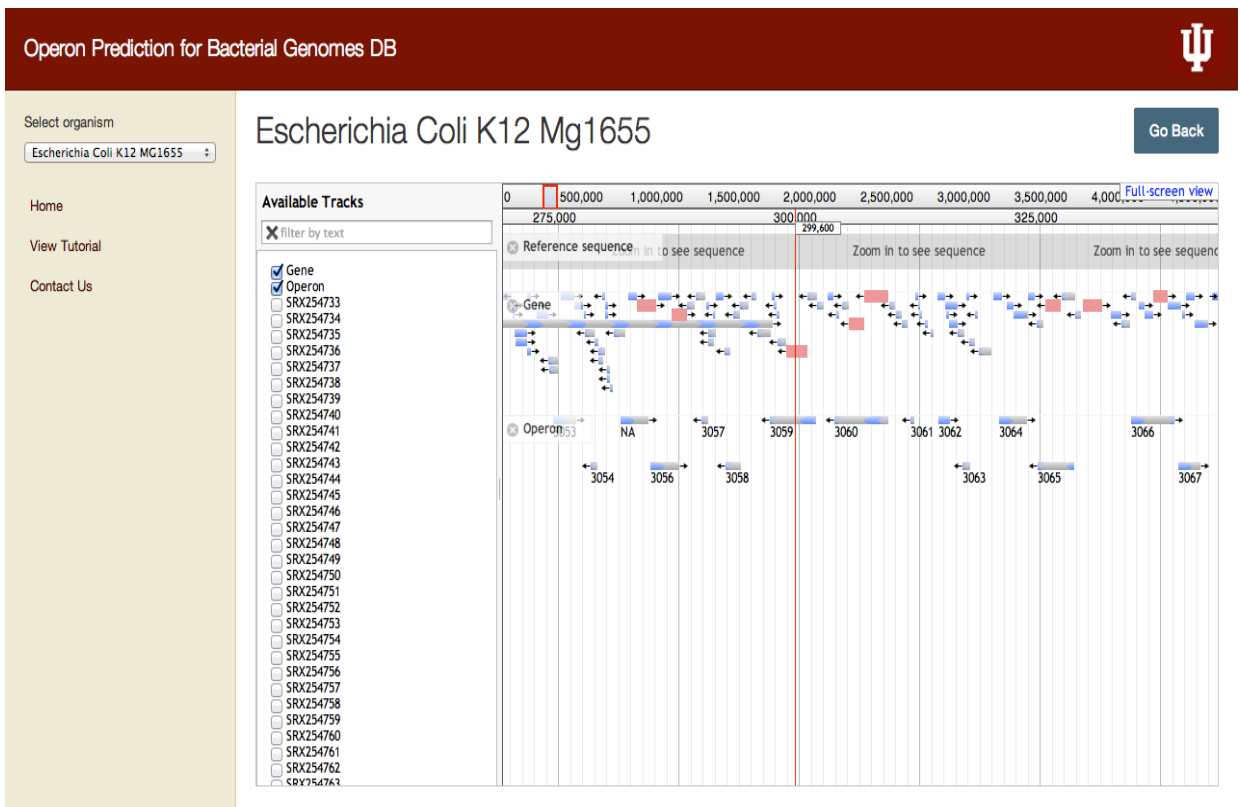
FASTA

```
>gi|556503834|ref|NC_000913.3|
gi|556503834|ref|NC_000913.3|:303719..309250 (- strand)
class=3060 length=5532
CGCGTCTGACATAAGTGGCGCAGCAATAGGTGGGGTGATTATTCGCAGGCCTTCAGTCAGGC
GCTTCAGGACGGCATGAGCGTCCCGCTTATATTCATCTGCCGGTAGCCAGGGTCCGCAGGAC
GATCAGGCAATCGGCAGCGCTTTTATCTGGCTGGGAGATGGACACTACGCATCCGGAAATAC
AGCTGGAAGAGAGTGAAGATAACCCAGTGTGACGGAACAACCTGCACAGCAGCTGATGGCTCT
```



**Figure 2. Screenshot showing the selection of an operon in the operons track. Highlighted is the ecpBCDE operon in Escherichia coli K12 genome encoding for the membrane and fimbria formation proteins. This view provides the name (database generated ID), position, type and length of the operon. It also gives information such as the number of genes present in the operon and sequence of the region for the selected operon.**

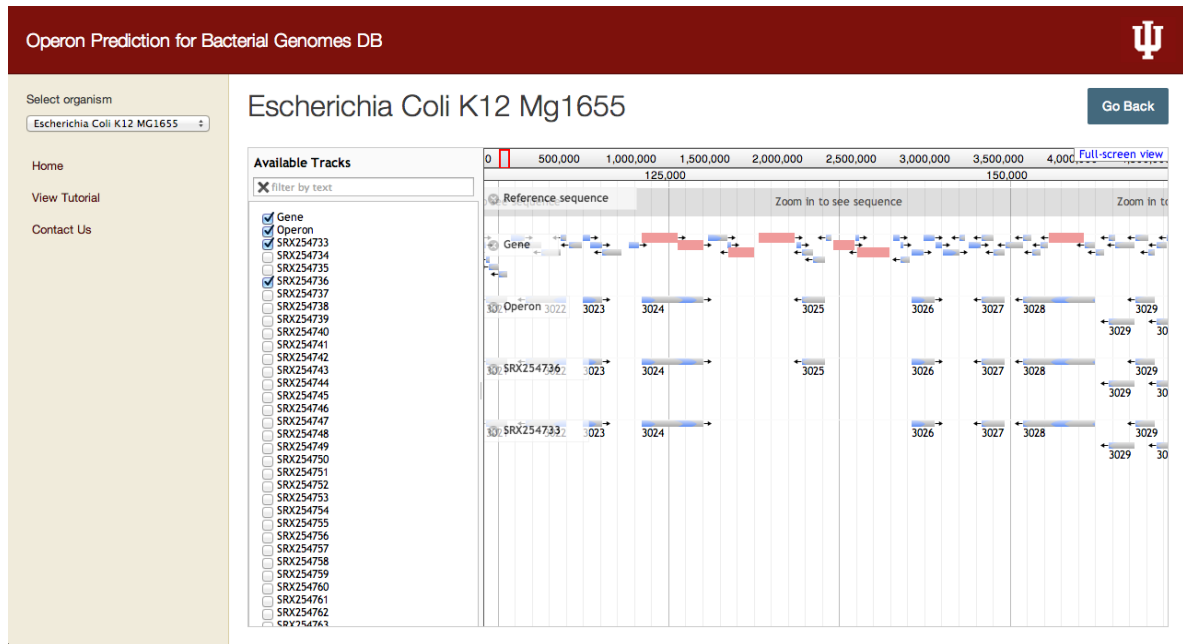




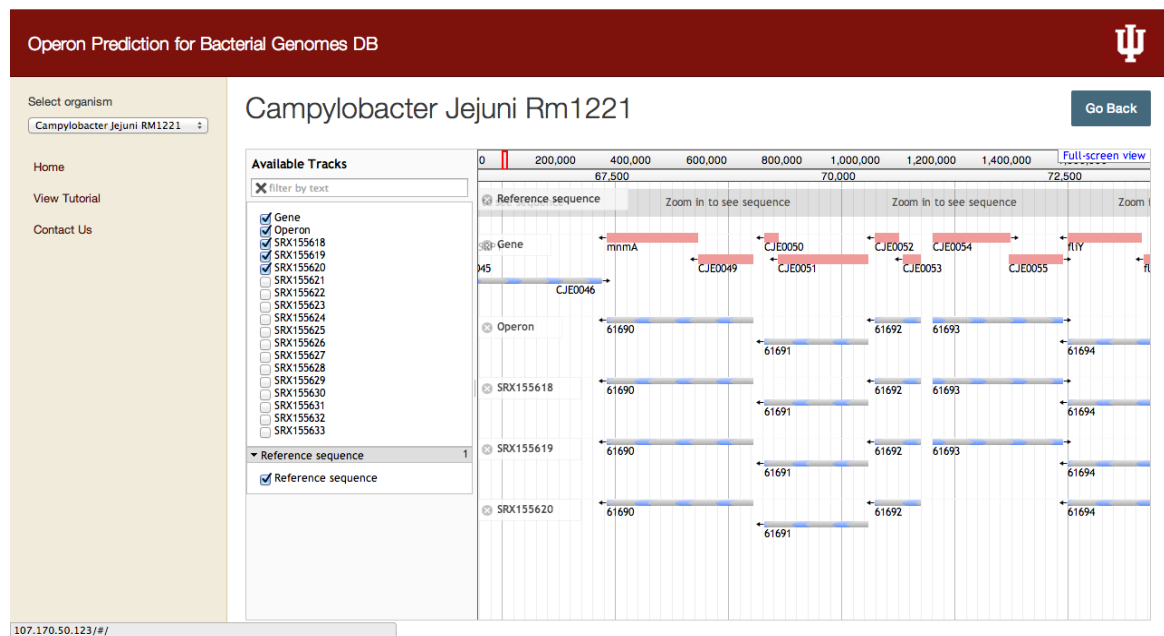
**Figure 3. Snapshot of the jBrowse visualization showing the ensemble of all the operons predicted for a bacterial organism.**

User can select the reference sequence; genes present in the organism, operons predicted from all the datasets as well as select the individual dataset to get the operons predicted for a particular experimental condition.

(A)

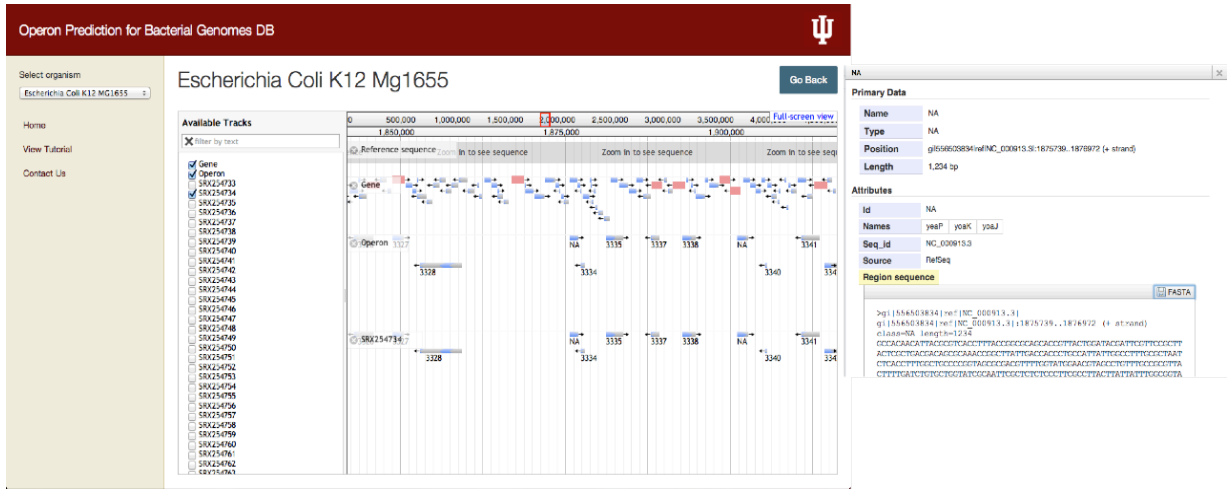


(B)



**Figure 4. Presence and absence of operons for different experimental conditions in two different bacterial genomes.**

(A) for *Escherichia coli* K12 MG1655 we have displayed the information for the missing operon '3025' in one of the experimental conditions - SRX254733. (B) Another example is from *Campylobacter jejuni* Rm1221 where we have displayed the information for the condition SRX155620 with missing operon '61693'.



**Figure 5. Operonic view showing a newly identified operon (yeaP-yoak-yoaj) in the genome of *Escherichia coli* K12. In operomeDB, newly identified operons compared to other databases such as DOOR are marked as 'NA' and user can further click on these to get the relevant information.**

## Chapter 3 Genomic and mechanistic Insights of convergent transcription in bacterial genomes

### 3.1 Introduction

In the process of gene expression, transcription is the first step in which DNA is converted into RNA. A complementary antiparallel strand is produced when the DNA is read by RNA polymerase. When two convergent promoters locate on a DNA strand, it is called *convergent transcription* (Crampton, et al.). Convergent transcription is the simultaneous induction of the sense and antisense transcription through two different opposing promoters (Lin, et al.). Sense and antisense transcripts frequently co-occur in genomic proximity in prokaryotic and eukaryotic organisms, so convergent transcription provides a model in the process of gene expression control. Many studies have been conducted to document the role of convergent transcription, providing evidence for its functionality and gene expression control in various organisms. Convergent gene transcription also allows an understanding of the biology and the process of transcriptional gene silencing by RNA interference mechanisms due to pervasive overlapping transcripts produced in most genomes (Gullerova and Proudfoot).

Chatterjee et al. described the role of convergent transcription in acting as a bistable switch in the process of antibiotic synthesis in *Streptomyces coelicolor* (Chatterjee, et al.). The authors showed precise expression control via antisense regulation acting as a bistable switch and thus stabilizing specific concentrations in the production of antibiotics (Chatterjee, et al.). In another study by the same authors, convergent transcription was postulated to act as a bistable switch across other species, suggesting that the mechanism of coupling RNA polymerase (RNAP) collision and

antisense interaction have an important regulatory role in gene expression in bacterial systems (Chatterjee, et al.). In particular, the authors showed that convergent transcription in the prgX and prgQ operon in *Enterococcus faecialis* enable the system with several properties of a genetic switch with premature termination of elongating transcripts due to collisions between RNAPs transcribing from different directions and an antisense regulation between the resulting complementary counter-transcripts (Chatterjee, et al.). Studies have also shown that head-on collision of RNAPs can hinder transcription in eukaryotes (Hobson, et al.). In yeast, it was shown that RNAP collision stops transcription when there is head-to-head collision, further demonstrating the role of convergent transcription in the process of gene regulation (Hobson, et al.).

Studies on viral RNA suggest that there is a suppression of UGA by tRNA. UGA is the stop codon that gets suppressed by tRNA unlike UAG and UAA. The authors also discussed the mismatch of base position in the transcription process (Urban, et al.). In a different study, it is shown that in bacteria, TGA is the leading stop codon, which is probably due to the abundance of GC content in their genome (Wong, et al.). Another study suggests that TGA has higher adaptability for biological mutations in bacteria than do TAA and TAG codons and that is because the frequency and fitness for TAA and TAG are dependent upon their GC content (Povolotskaya, et al.). The probability of use of premature stop codons (PSC's) content truly depends on the GC content of the bacterial genome, but not all bacteria contain that significant amount of GC content, which could possibly make the high usage of UGA universal for all bacterial species. All conserved genes are necessary to define on the basis of their homologous and orthologous relationships to obtain information about genome sequences. Clusters of orthologous groups (COGs) of proteins represent a

phylogenetic classification encoded in complete genomes, and this classification helps in providing a functional insight into various genomes (Tatusov, et al.).

For this project, we used 2,168 bacterial genomes to predict their functionality, lifestyle, distribution and conservation for convergent overlapping gene pairs across transcription. For the study, we first paired the genes as FF or RR, FR and RF; from this we selected the gene pairs with FR strands, which are responsible for convergent transcription. Gaussian density distribution was used to predict the highest density distribution point and the inter-genic distance for each genome having COGPs. This led us to find the conservation pattern across all genomes, where most of the convergent transcription occurs. To predict functionality, we used COGs and mapped them to the FR strand to anticipate their functionality across convergent gene pairs.

### 3.2 Material and Methods

#### **Genome size distribution across bacterial genomes**

To identify the distribution of genome size (Mbp) across COGPs, we calculated the proportion of 3' to 3' adjacent gene pairs for each organism. We then mapped the proportion of COGPs to genome size and plotted a distribution for different bacterial groups, giving us an understanding of the significance of genome size across COGPs.

### **Bacterial lifestyle distribution across COGPs**

We studied the distribution of proportion of COGPs for different bacterial lifestyles across bacteria and their effect on COGPs. For this, we plotted a multi-panel boxplot for COGPs with different lifestyles. We considered four lifestyles for bacteria endospore formation, oxygen requirement, habitat and temperature to study their effect on COGPs. We used a Wilcoxon test to determine the significance and performed ANOVA and multiple regression to study the individual and combined effects of each factor on COGPs across bacterial population.

### **Density distribution of COGPs**

To determine the intergenic distance density distribution of COGPs across bacterial genomes, we plotted the intergenic distance for COGPs for all 2,168 bacterial genomes available at NCBI's RefSeq database (Tatusova, et al., 2014). Using the R programming environment, we plotted the density distribution for intergenic distance to determine the highest density distribution point for COGPs across the intergenic distance range of 0 to -50 to 0 bp.

### **Identify COGPs conserved overlaps**

To identify the overlaps conserved across genomes, a matrix was generated for all convergent gene pairs for a particular genome across all other genomes. The proportion of COGPs was calculated by taking a count of the negative gene pairs to total number of positive and negative counts. A graph was generated with the proportion of negative convergent gene pairs and the inter-genic distance for different genomes.

### **Mapping of COGs to Convergent gene pairs**

COGs are groups of clusters found to be orthologous across at least three lineages. Each COG has a specific functional description; it may also have one or more general categories. Each COG represents one of 23 different functional categories. To predict the functional enrichment of convergent gene pairs, we mapped each COG and its functional category to convergent gene pairs. This allowed us to determine the symmetry of the COGs and their functional categories. For each functional category, we mapped gene pairs and calculated their hypergeometric test using the `dhyper` function in R programming language (Rivals, et al.). Further, we generated a heat map for 23 functional categories to learn the significant functional categories that might help in the regulation of gene functioning.

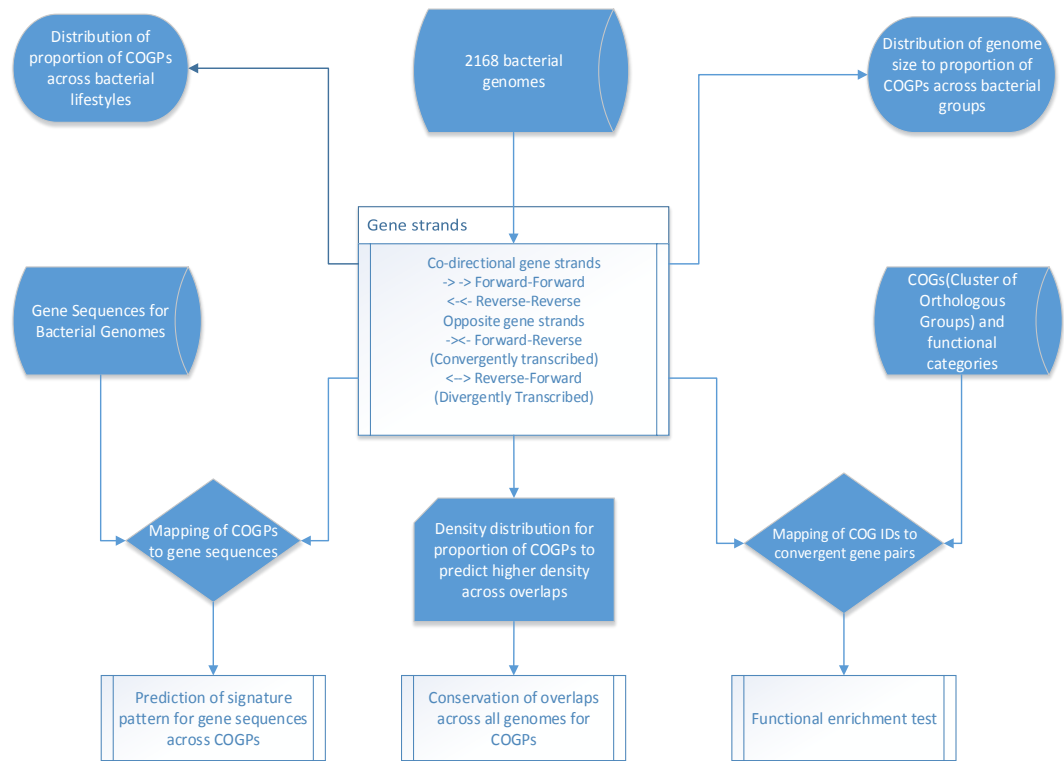
### **Mapping of Gene Sequence to COGPs**

Gene sequences of all bacterial genomes were mapped to COGPs to further investigate various signature patterns of overlapping gene pairs. Mapping was done to enquire about the various functions and processes related to COGPs. We used WebLogo to illustrate sequence pattern of COGPs with intergenic distance overlaps of -4 and -11 respectively (Crooks, et al.)

## **3.3 Results and discussion**

The workflow and methodology we used to predict the mechanistic insights in bacterial genomes for convergent transcription is discussed in Figure 6. For this study, we used bacterial genomes, gene replicon information and a COGs database to learn various functionalities and distribution of convergent transcription in bacteria.





Methodology to identify mechanistic insights in bacterial genomes

**Figure 6. Flowchart for methodology and processing of bacterial genome data.** Flowchart describes the methodology and the process we used to determine the genomic and mechanistic insights of the bacterial genomes for convergent transcription

### **Bacterial genomes of different groups show significant correlations with genome size for COGPs**

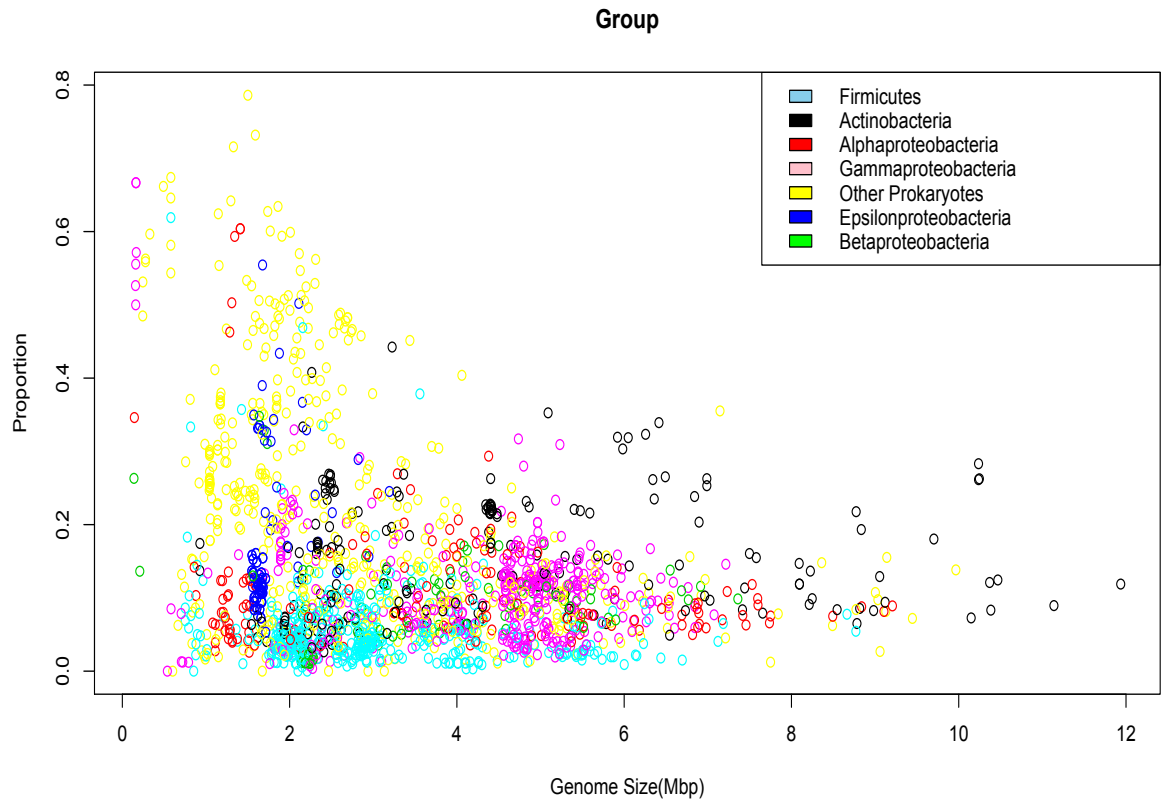
Bacterial genomes show a negative correlation between genome size (Mbp) and their proportion of convergent overlapping gene pairs (COGPs) ( $R = -0.1334291$ ;  $p = 1.634038e-09$ ). A previous study suggested that overlapping gene pairs are related to gene expression regulation and minimization of genomes (Johnson and Chisholm). In agreement, we found that the proportion of COGPs tends to decrease with increase in genome size (Mbp) across various groups of bacterial genomes

(Figure 7). The distribution patterns across phyla showed that Actinobacteria have larger genome sizes with lower proportions of COGPs. Gamma-Proteobacteria showed higher proportions of COGPs with smaller genome sizes.

### **Significant and wide distribution of COGPs across bacteria with different lifestyles**

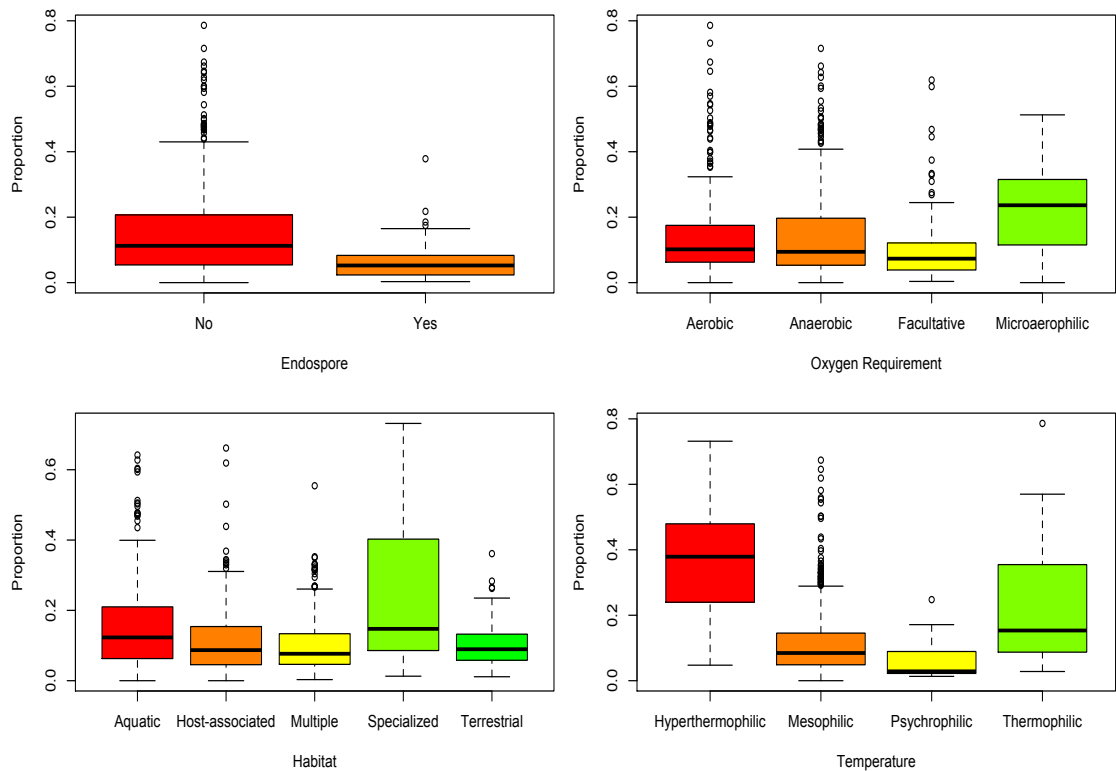
To better understand the distribution and roles of COGPs, we studied different lifestyles of bacteria, such as oxygen requirement, endospore formation, habitat, temperature range, shape, motility, genome size and pathogenicity. We plotted the proportion of COGPs with different lifestyles (Figure 8).

In agreement with previous studies (Nicholson, et al.), endospore formation is present in just a few of the bacterial organisms examined. Bacteria that do not form endospores have a significantly higher proportion of COGPs (Table 1; Figure 8A). According to our results, bacteria growing with in conditions with very low levels of oxygen requirement, microaerophilic, tend to have higher proportions of COGPs (Table 1; Figure 8B).



**Figure 7. Distribution of proportion of COGPs across genome size for bacterial genome groups**

Proportions of COGPs across genome size (Mbp) represent that the higher the proportion of COGPs, the less their genome size across bacterial groups. Some of the gammaproteobacteria have high proportion and less genome size. Therefore, there is a pattern across genomes in which the proportion of COGPs decreases with the increase in genome size.



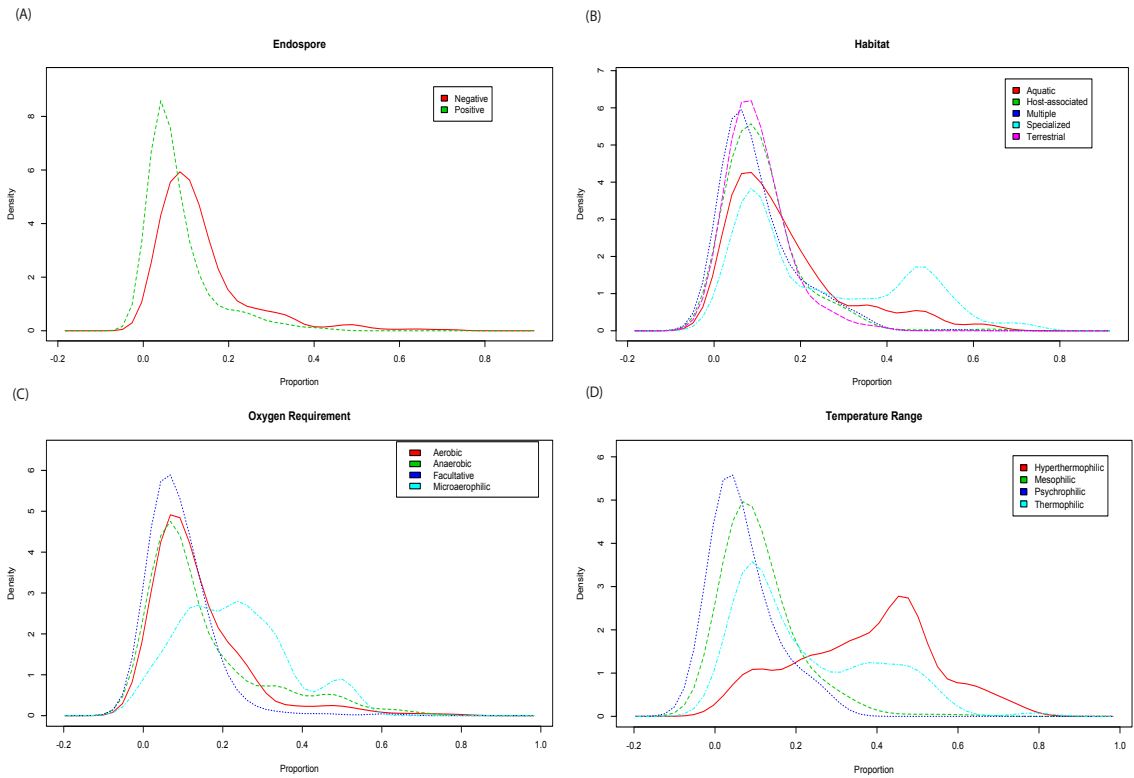
**Figure 8. Comparison of proportion of COGPs to different lifestyle for bacterial genomes**

Different lifestyles of bacteria include endospore formation, oxygen requirement, habitat and temperature range. Proportion of COGPs for bacterial genomes was plotted with different lifestyles to learn about its factors and the pattern of different conditions or factors across COGPs. (A) For endospore, the negative factor has a high proportion, which explains that endospore is not present for bacteria with high COGPs proportion. (B) Oxygen requirement explains the level of oxygen required by different bacteria to survive. Microaerophilic shows high abundance for COGPs, which means bacteria with high COGPs can thrive in a low level of oxygen. (C) Specialized and aquatic habitats have a high proportion of COGPs for bacterial genomes. The proportion of COGPs is higher in bacteria for specialized habitats than in other habitats. (D) Hyperthermophilic and thermophilic is the temperature range in which bacteria with the highest proportion of COGPs can survive.

Specialized and aquatic habitats are not completely understood (Sunagawa, et al.). We find that a specialized habitat has the highest proportion of COGPs followed by an aquatic habitat (Figure 8) with significant p-values (Table 1). We found that a specialized habitat has the highest proportion of COGPs followed by an aquatic

habitat (Table 1; Figure 8C). Bacteria that thrive in psychrophilic and mesophilic conditions have higher density distribution than thermophilic bacteria (Figure 9). However, thermophilic and hyperthermophilic bacteria have a higher proportion of COGPs (Figure 3D) with significant p-values (Table 1).

To estimate the relationship between proportion of COGPs and other lifestyles, analysis of variance (ANOVA) was performed for individual and combined lifestyles. Many of these were significant (Table 2). We performed a multiple regression analysis (R-squared = 0.5986; p-value < 2.2e-16) of proportion of COGPs and different lifestyles (Table 5). From the analysis, we confirmed that bacteria with no endospore formation, high level of GC content, thermophilic nature, and living in specialized habitats with low oxygen levels show a high proportion of COGPs.



**Figure 9. Density distribution plots for various lifestyles on the basis of their proportion.**

**Table 1. Detailed p-value calculated using paired Wilcoxon test for the proportion of COGPs across bacterial genomes for different lifestyles**

Lifestyle	Paired factors	p-value
Endospore	Positive and Negative	< 2.2e-16
Oxygen Requirement	Aerobic and Facultative	7.565e-12
	Aerobic and Microaerophilic	3.832e-06
	Anaerobic and Microaerophilic	4.329e-05
	Anaerobic and Facultative	1.092e-06

	Microaerophilic and Facultative	1.326e-11
Habitat	Multiple and Specialized	< 2.2e-16
	Multiple and Aquatic	4.758e-09
	Aquatic and Terrestrial	0.0003638
	Aquatic and Host-Associated	1.189e-05
	Aquatic and Specialized	0.001327
	Terrestrial and Specialized	1.501e-08
	Specialized and Host-associated	4.587e-13
Temperature Range	Mesophilic and Thermophilic	8.238e-13
	Mesophilic and Hyperthermophilic	< 2.2e-16
	Thermophilic and Psychrophilic	9.231e-05
	Thermophilic and Hyperthermophilic	4.583e-07
	Psychrophilic and Hyperthermophilic	4.779e-07

### **COGPs across bacterial genomes tend to exhibit shorter overlaps**

The density distribution of COGPs on the basis of intergenic distance across bacterial genomes can provide insight into the distribution and their overlaps. We performed a Gaussian density distribution analysis of COGPs intergenic distances for 2,168 bacterial genomes. Figure. 10 shows the density distribution for six organisms: *Escherichia Coli*, *Bacillus Subtilis*, *Helicobacter Pylori*, *Synechocystis*, *Shigella Boydii* and *Synechococcus* sp. We observed that COGPs tend to exhibit high densities at intergenic distances corresponding to short overlaps (approx. -4 to 0 bp). For better understanding of the overlapping region, the density distribution was plotted for all bacterial genomes by selecting the range of intergenic distance -50

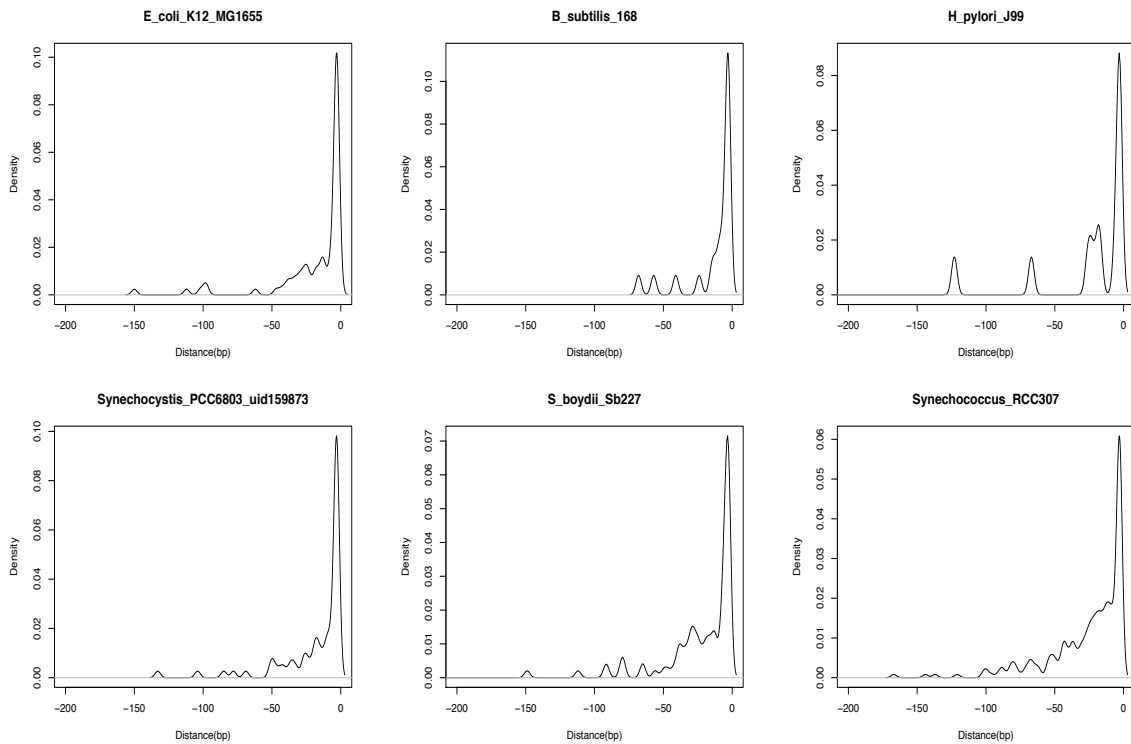
to 0 (Figure S1). From the data we can conclude that COGPs across bacterial genomes tend to have short overlaps with a prevalent intergenic distance of -4.

Table 2. P-values for individual and combined factors for different lifestyles calculated using anova. Proportion of COGPs was taken as a response variable to other lifestyles, and p-values were calculated for each.

Individual factors	p-value for Individual factors	Combined factors	p-value for Combined factors
GC content	<2e-16	Genome size: temperature	1.98e-12
		Genome size: shape: oxygen	3.00e-11
Temperature	<2e-16	GC content: oxygen requirement	1.69e-09
		Shape: oxygen requirement	8.23e-09
Habitat	3.25e-15	GC content: shape	4.78e-08
		Endospore: oxygen requirement: habitat	6.02e-08
Oxygen Requirement	7.65e-12	Genome size: habitat	6.84e-08
		Genome size: GC content: temperature	1.55e-07
Genome size	2.54e-10	Genome size: oxygen requirement	4.16e-07
		Endospore: oxygen requirement	1.06e-05
Motility	2.51e-08	Genome size: shape	1.45e-05
		Oxygen requirement: habitat	2.23e-05
Endospore	1.97e06	Genome size: shape: temperature	4.56e-05
		GC content: shape: oxygen requirement	5.29e-05
Shape	2.81e-06	Genome size: GC content	0.000154
		Shape: habitat	0.000321
Pathogenicity	0.00194	GC content: oxygen requirement: habitat	3.56e-040.00038



		GC content: endospore Genome size: habitat: temperature	7 1.39e-03
--	--	---	---------------



**Figure 10. Density distribution of bacterial genomes for COGPs**

Gaussian density distribution across COGPs was calculated for each bacterial genome to learn about its overlapping pattern. It showed that most COGPs with shorter intergenic distance have high density distribution for bacterial genomes.

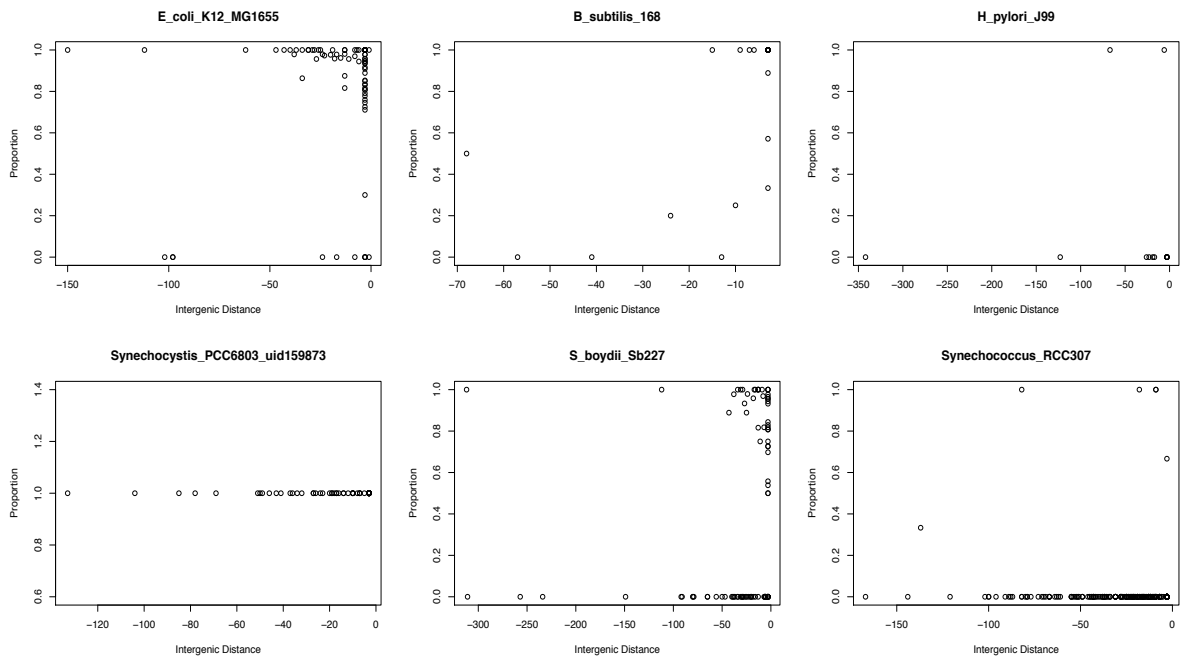
### **COGPs with shorter overlaps tend to be more conserved**

The conservation pattern across microbial genomes has also been studied previously, depicting that conservation of overlapping genes is within small overlapping sequence regions (Johnson and Chisholm). For studying the conservation pattern of COGPs across bacterial genomes, we mapped the genes of a particular genome to their orthologous genes in all other genomes to study their conservation patterns. We generated a plot of conservation of COGPs organized by the intergenic distance of the reference COGPs (Figure 11). Overall, these results suggested that COGPs with shorter overlaps are the most conserved across bacterial genomes, with prevalence, again, of the -4 bp intergenic distance overlap. This analysis also explains the high conservation and proportion of COGPs that increase with the decrease in inter-genic distance.

### **Functional enrichment of convergent gene pairs using COGs**

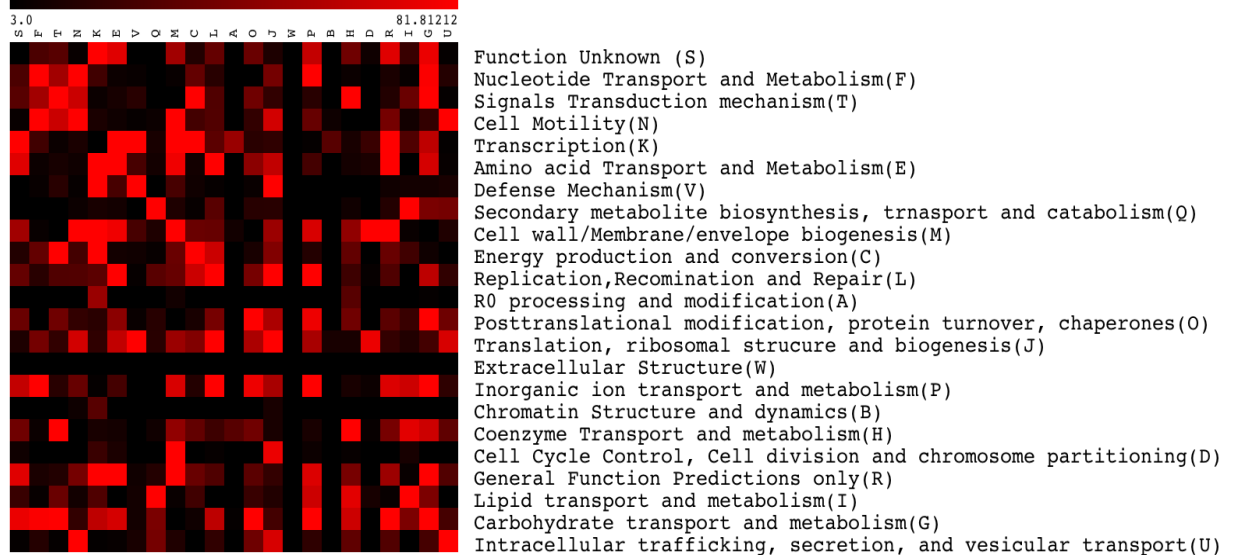
To investigate the possibility of a bias in functions performed by genes found in COGPs, we mapped genes in COGPs to their Clusters of Orthologous Groups (COG) IDs and their corresponding 23 COG functional categories (Tatusov, et al.). We constructed a heatmap (Figure 12) showing the enrichment of COG category pairs found in COGPs (using hypergeometric test (Rivals, et al.),  $-\log$  of p-value at 1% FDR) Many significant functional categories those that regulate the functioning of many bacterial genes were identified from the heatmap. Most of the functional categories have a significant enrichment in cell motility, cellular processes and signaling, metabolism, processing and storage of information. Most of the functions, which are significant, interact through COGPS pairs are 1) translational, ribosomal structure and biogenesis (J) with defense mechanism (V), 2) nucleotide transport

and metabolism (F) with inorganic transport and metabolism, 3) signal transduction mechanism (T) with co-enzyme transport and metabolism (H), 4) replication, recombination and repair (L) with translational, ribosomal structure and biogenesis (J), 5) transcription (K) with amino acid transport and metabolism (E) and defense mechanism (V).



**Figure 11. Conservation of COGPs across shorter overlaps for all genomes**

COGPs are conserved across shorter overlaps for all genomes. One genome was mapped across all genomes to determine the distribution across COGPs and its intergenic distance. Most of the genome shows conservation of COGPs for shorter overlaps across bacterial genomes.



### Figure 12. Functional enrichment of convergent gene pairs

Mapping was done for convergent gene pairs using COG IDs and functional category. For this Hypergeometric distribution was done to calculate the p-value for convergent gene pairs, then the calculated p-values were adjusted using 1% FDR and, after that, a negative log of p-value was calculated to obtain the significant functional categories for convergent gene pairs.

### Stop codon prediction from shorter overlapping COGPs

To further investigate the shorter overlaps that are conserved, we mapped gene sequences to COGPs. We observed that the process of convergent transcription contains stop codons, with inter-genic distances  $-4$  and  $-11$  bp. We designed a WebLogo to represent the overlapping stop codons (Fig.13) (Crooks, et al.).

We further analyzed COGPs with intergenic distance overlaps of  $-4$  and less than  $-4$  by mapping them to gene sequences and finding that various stop codons

associated with COGPs have 1,2 and 3 base position mismatches (Table 3a and b). Previous studies have suggested that codons that reduce the folding of mRNA at the starting of translation process are favored in bacteria (Bentele, et al.).

We also calculated the proportion of stop codon for COGPs with intergenic distance  $-4$  and less than  $-4$  (Table 4). For obvious compatibility issues, the intergenic distance overlap of  $-4$  contains only two stop codons, which are TAG and TAA (48% and 52% respectively). COGPs at intergenic distance of less than  $-4$  contain all three stop codons: TAG, TGA and TAA (24%, 46% and 30% respectively). Previously published studies suggested that TGA was present in overlapping gene sequences with a high frequency of GC content, while TAG and TAA were prevalent in genes with lower GC frequency, despite the fact that TAG and TGA have the same composition, suggesting TGA has high adaptability than do TAG and TAA for biological mutations (Povolotskaya, et al.; Wong, et al.). As previously discussed in the results, GC content is significant for COGPs and we also know that TGA has a higher percentage in gene sequence with intergenic distance less than  $-4$ . This confirms that TGA occurs more frequently in high GC content COGPs.

Table 3. Percentage and mismatch of stop codons for COGPs when mapped to gene sequence (A) determines the COGPs for inter-genic distance less than  $<-4$  (B) It determines the COGPs for inter-genic distance  $-4$

(A)

---

Gene Pairs	Percentage	No. Of base Mismatch
TGA AGT	25.4%	1 base mismatch
TGA AAT	10.3%	1 base mismatch
TAA AGT	9.9%	1 base mismatch
TAA AAT	13.3%	1 base mismatch
TGA GAT	11%	2 base mismatch
TAA GAT	6.7%	2 base mismatch
TAG AAT	6.5%	2 base mismatch
TAG AGT	10.8%	2 base mismatch
TAG GAT	6%	3 base mismatch

---

(B)

---

Gene Pairs	Percentage	No. Of base Mismatch
TAA AAT	33%	1 base mismatch
TAG AAT	19%	2 base mismatch
TAA GAT	19%	2 base mismatch
TAG GAT	29%	3 base mismatch

---

Table 4. Percentage of each stop codon across COGPs (A) for intergenic distance -4, (B) for intergenic distance less than <- 4

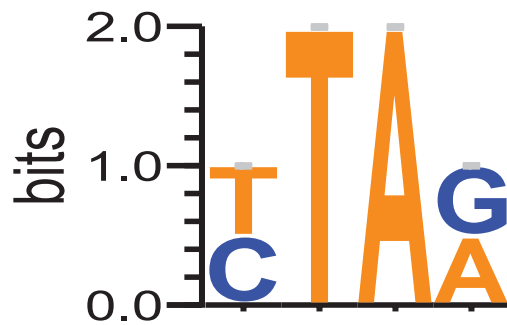
(A)

Stop Codon	Percentage
TAG	48%
TAA	52%

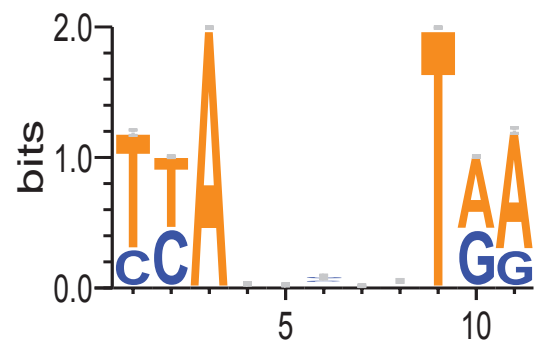
(B)

Stop Codon	Percentage
TAG	24%
TGA	46%
TAA	30%

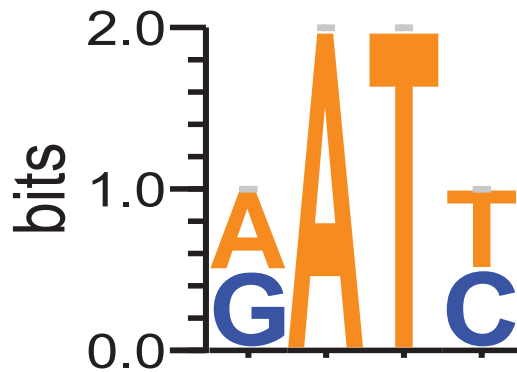
(A) Forward Strand



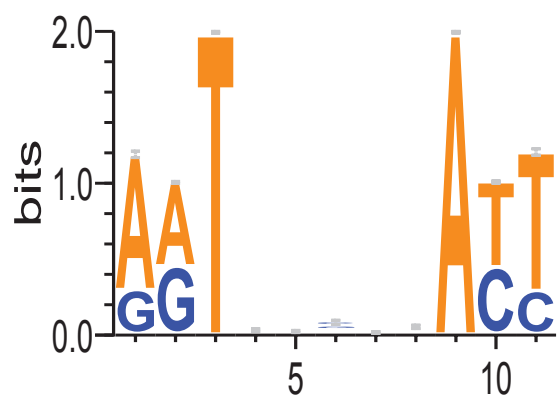
(B) Forward Strand



Reverse Strand



Reverse Strand



**Figure 13. Sequence logo representation for gene sequence.** A sequence logo was developed for COGPs that have intergenic distance (A) -4 and (B) -11. The logo describes the forward strand and reverse strand for intergenic distance. This figure informed us about the pattern occurring across the COGPs, which shows the presence of stop codons.



Table 5. This table contains the significant p-values for individual factors that are significant for a particular lifestyle. Adjusted R-squared value and p-value were also calculated using multiple regression, which shows all the factors significant for bacteria with COGPs.

Factors for Lifestyles	p-value for individual factors	Adjusted R-squared value and p-value for all factors
GC content	4.66e-07	Adjusted R-squared value= 0.5986  p-value for all = <2.2e-16
Endospore [Yes]	1.16e-07	
Oxygen Requirement [Anaerobic]	0.00728	
Temperature [Thermophilic]	1.88e-10	
Temperature [Mespohilic]	<2e-16	
Pathogenicity [human, animal, insect]	7.87e-05	
Pathogenicity [plant]	0.00905	

## Chapter 4 Conclusion

Bacteria are considered simpler organisms than are humans, and they are easy to study; however it is fairly clear that gene regulation in bacteria is extremely efficient. As bacterial organisms are highly organized, they can thrive in different environmental conditions and can respond to environment changes by adapting themselves accordingly.

In our study, genomics and mechanistic insights of convergent transcription in bacterial organisms, we have discussed the various processes that help in the regulation of genes and how they adapt to different environments with relation to convergent transcription. In this study, we explored the role of various factors contributing to convergent overlapping transcription in bacterial genomes. Our analysis showed that the proportion of convergent overlapping gene pairs (COGPs) in a genome is affected due to endospore formation, bacterial habitat, oxygen requirement, GC content and temperature range. In particular, we showed that bacterial genomes thriving in specialized habitats, such as thermophiles, exhibit a high proportion of COGPs. Our results also showed that the density distribution of COGPs across the genomes is high for shorter overlaps with increased conservation of distances for decreasing overlaps. Our study also revealed that COGPs frequently contain stop codon overlaps with the middle base position exhibiting mismatches between complementary strands. Functional analysis using COGs annotations suggested that cell motility, cell metabolism, storage and cell signaling are enriched among COGPs, suggesting their role in process that go beyond regulation. Thereby, our analysis provided genomic insights into this unappreciated regulatory

phenomenon, allowing a refined understanding of their contribution to bacterial phenotypes.

Characterizing operon structures in a genome is one of the first and fundamental steps towards improving our understanding on transcriptional regulation in bacterial genomes. OperomeDB represents one of the first attempts to provide a comprehensive resource for operon structures in microbial genomes based on RNA-sequencing data, providing a one stop portal for understanding the genome organization in the context of transcriptional regulation in a condition-specific manner. OperomeDB as a database, should not only aid experimental groups working on transcriptome analysis of specific organisms but also enable studies related to computational and comparative operomics.

In our study each SRA ID for which the operon prediction was performed corresponds to a different condition or perturbation to the cell in which RNA was sequenced to quantitate the expression levels of genes. Therefore, this database will not only be helpful for researchers to browse through each condition and analyze operons predicted for that particular condition but also to add their own new RNA-seq datasets corresponding to their experiments to uncover novel operon signatures specific to their condition of interest. Researchers can also compare operons predicted in our database with other databases under various conditions. Comparing operons under experimental and normal conditions will provide insight into the mechanism and effect of the particular condition on bacterial regulation at specific genomic loci. In the future, we will add more bacterial organisms with RNA-seq datasets to our database and we will also increase the number of datasets/conditions for already existing bacterial organisms in our database.

## Chapter 5 Future Work

In our future work, we would like to increase the number of bacterial organism in our database for operon prediction. We also desire to include other scientific data from different sources to the database. We plan on making the web interface more effective and user friendly; so that this tool can be of help to show bigger, a better picture of genomes.

## Chapter 6 References

(Babitzke, 2004; Bentele, et al., 2013; Browning and Busby, 2004; Chatterjee, et al., 2011; Chatterjee, et al., 2011; Chivian, et al., 2013; Crampton, et al., 2006; Crooks, et al., 2004; Dugar, et al., 2013; Fimlaid, et al., 2013; Flicek, et al., 2014; Goldman, et al., 2013; Guell, et al., 2009; Gullerova and Proudfoot, 2012; Haiser, et al., 2013; Hobson, et al., 2012; Johnson and Chisholm, 2004; Kuhn, et al., 2013; Lin, et al., 2010; Mao, et al., 2009; McClure, et al., 2013; Nicholson, et al., 2000; Pertea, et al., 2009; Povolotskaya, et al., 2012; Rivals, et al., 2007; Ruffing, 2013; Salgado, et al., 2013; Sallet, et al., 2013; Scheffers and Errington, 2004; Skinner, et al., 2009; Stringer, et al., 2014; Stulke, 2002; Sunagawa, et al., 2010; Taboada, et al., 2012; Tatusov, et al., 2000; Tatusov, et al., 1997; Thorvaldsdottir, et al., 2013; Uplekar, et al., 2013; Urban, et al., 1996; Wang, et al., 2011; Winkler and Breaker, 2005; Wong, et al., 2008)

Babitzke, P. (2004) Regulation of transcription attenuation and translation initiation by allosteric control of an RNA-binding protein: the *Bacillus subtilis* TRAP protein, *Current opinion in microbiology*, **7**, 132-139.

Bentele, K., et al. (2013) Efficient translation initiation dictates codon usage at gene start, *Molecular systems biology*, **9**, 675.

Browning, D.F. and Busby, S.J. (2004) The regulation of bacterial transcription initiation, *Nature reviews. Microbiology*, **2**, 57-65.

Chatterjee, A., et al. (2011) Convergent transcription in the butyrolactone regulon in *Streptomyces coelicolor* confers a bistable genetic switch for antibiotic biosynthesis, *PLoS one*, **6**, e21974.

Chatterjee, A., et al. (2011) Convergent transcription confers a bistable switch in *Enterococcus faecalis* conjugation, *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 9721-9726.

Chivian, D., et al. (2013) MetaMicrobesOnline: phylogenomic analysis of microbial communities, *Nucleic acids research*, **41**, D648-654.

Crampton, N., et al. (2006) Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy, *Nucleic acids research*, **34**, 5416-5425.

Crooks, G.E., *et al.* (2004) WebLogo: a sequence logo generator, *Genome research*, **14**, 1188-1190.

Dandekar, T., *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact, *Trends in biochemical sciences*, **23**, 324-328.

Dugar, G., *et al.* (2013) High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates, *PLoS genetics*, **9**, e1003495.

Ermolaeva, M.D., White, O. and Salzberg, S.L. (2001) Prediction of operons in microbial genomes, *Nucleic acids research*, **29**, 1216-1221.

Fimlaid, K.A., *et al.* (2013) Global analysis of the sporulation pathway of *Clostridium difficile*, *PLoS genetics*, **9**, e1003660.

Flicek, P., *et al.* (2014) Ensembl 2014, *Nucleic acids research*, **42**, D749-755.

Goldman, M., *et al.* (2013) The UCSC Cancer Genomics Browser: update 2013, *Nucleic acids research*, **41**, D949-954.

Guell, M., *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium, *Science (New York, N.Y.)*, **326**, 1268-1271.

Gullerova, M. and Proudfoot, N.J. (2012) Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells, *Nature structural & molecular biology*, **19**, 1193-1201.

Haiser, H.J., *et al.* (2013) Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*, *Science (New York, N.Y.)*, **341**, 295-298.

Hobson, D.J., *et al.* (2012) RNA polymerase II collision interrupts convergent transcription, *Molecular cell*, **48**, 365-374.

Jacob, F., *et al.* (1960) [Operon: a group of genes with the expression coordinated by an operator.], *C R Hebd Seances Acad Sci*, **250**, 1727-1729.

Jacob, F., *et al.* (2005) [The operon: a group of genes with expression coordinated by an operator. C.R.Acad. Sci. Paris 250 (1960) 1727-1729], *C R Biol*, **328**, 514-520.

Janga, S.C., Collado-Vides, J. and Moreno-Hagelsieb, G. (2005) Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons, *Nucleic acids research*, **33**, 2521-2530.

Johnson, Z.I. and Chisholm, S.W. (2004) Properties of overlapping genes are conserved across microbial genomes, *Genome research*, **14**, 2268-2272.

Kodama, Y., Shumway, M. and Leinonen, R. (2012) The Sequence Read Archive: explosive growth of sequencing data, *Nucleic acids research*, **40**, D54-56.

Kuhn, R.M., Haussler, D. and Kent, W.J. (2013) The UCSC genome browser and associated tools, *Briefings in bioinformatics*, **14**, 144-161.

Lathe, W.C., 3rd, Snel, B. and Bork, P. (2000) Gene context conservation of a higher order than operons, *Trends in biochemical sciences*, **25**, 474-479.

Lin, Y., *et al.* (2010) Convergent transcription through a long CAG tract destabilizes repeats and induces apoptosis, *Molecular and cellular biology*, **30**, 4435-4451.

Mao, F., *et al.* (2009) DOOR: a database for prokaryotic operons, *Nucleic acids research*, **37**, D459-463.

McClure, R., *et al.* (2013) Computational analysis of bacterial RNA-Seq data, *Nucleic acids research*, **41**, e140.

Nicholson, W.L., *et al.* (2000) Resistance of *Bacillus* endospores to extreme terrestrial and extraterrestrial environments, *Microbiology and molecular biology reviews : MMBR*, **64**, 548-572.

Overbeek, R., *et al.* (1999) The use of gene clusters to infer functional coupling, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2896-2901.

Pertea, M., *et al.* (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes, *Nucleic acids research*, **37**, D479-482.

Povolotskaya, I.S., *et al.* (2012) Stop codons in bacteria are not selectively equivalent, *Biology direct*, **7**, 30.

Rivals, I., *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test?, *Bioinformatics (Oxford, England)*, **23**, 401-407.

Ruffing, A.M. (2013) RNA-Seq analysis and targeted mutagenesis for improved free fatty acid production in an engineered cyanobacterium, *Biotechnology for biofuels*, **6**, 113.

Salgado, H., *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more, *Nucleic acids research*, **41**, D203-213.

Sallet, E., *et al.* (2013) Next-generation annotation of prokaryotic genomes with EuGene-P: application to *Sinorhizobium meliloti* 2011, *DNA research : an international journal for rapid publication of reports on genes and genomes*, **20**, 339-354.

Scheffers, D.J. and Errington, J. (2004) PBP1 is a component of the *Bacillus subtilis* cell division machinery, *Journal of bacteriology*, **186**, 5153-5156.

Sebahia, M., *et al.* (2006) The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome, *Nature genetics*, **38**, 779-786.

Seshasayee, A.S., *et al.* (2009) Principles of transcriptional regulation and evolution of the metabolic system in *E. coli*, *Genome research*, **19**, 79-91.

Skinner, M.E., *et al.* (2009) JBrowse: a next-generation genome browser, *Genome research*, **19**, 1630-1638.

Sorek, R. and Cossart, P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity, *Nature reviews. Genetics*, **11**, 9-16.

Stothard, P., *et al.* (2005) BacMap: an interactive picture atlas of annotated bacterial genomes, *Nucleic acids research*, **33**, D317-320.

Stringer, A.M., *et al.* (2014) Genome-scale analyses of *Escherichia coli* and *Salmonella enterica* AraC reveal noncanonical targets and an expanded core regulon, *Journal of bacteriology*, **196**, 660-671.

Stulke, J. (2002) Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures, *Archives of microbiology*, **177**, 433-440.

Sunagawa, S., Woodley, C.M. and Medina, M. (2010) Threatened corals provide underexplored microbial habitats, *PloS one*, **5**, e9554.

Taboada, B., *et al.* (2012) ProOpDB: Prokaryotic Operon DataBase, *Nucleic acids research*, **40**, D627-631.

Tatusov, R.L., *et al.* (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution, *Nucleic acids research*, **28**, 33-36.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science (New York, N.Y.)*, **278**, 631-637.

Tatusova, T., *et al.* (2014) RefSeq microbial genomes database: new representation and annotation strategy, *Nucleic acids research*, **42**, D553-559.

Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Briefings in bioinformatics*, **14**, 178-192.

Uplekar, S., *et al.* (2013) High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in *Mycobacterium tuberculosis*, *Nucleic acids research*, **41**, 961-977.

Urban, C., *et al.* (1996) UGA suppression by tRNACmCATrp occurs in diverse virus RNAs due to a limited influence of the codon context, *Nucleic acids research*, **24**, 3424-3430.

Wadler, C.S. and Vanderpool, C.K. (2009) Characterization of homologs of the small RNA SgrS reveals diversity in function, *Nucleic acids research*, **37**, 5477-5485.

Wang, Y., *et al.* (2011) Single-nucleotide resolution analysis of the transcriptome structure of *Clostridium beijerinckii* NCIMB 8052 using RNA-Seq, *BMC genomics*, **12**, 479.

- Westesson, O., Skinner, M. and Holmes, I. (2013) Visualizing next-generation sequencing data with JBrowse, *Briefings in bioinformatics*, **14**, 172-177.
- Winkler, W.C. and Breaker, R.R. (2005) Regulation of bacterial gene expression by riboswitches, *Annual review of microbiology*, **59**, 487-517.
- Wong, T.Y., *et al.* (2008) Role of premature stop codons in bacterial evolution, *Journal of bacteriology*, **190**, 6718-6725.



# KASHISH CHETAL

322 Canal Walk, Apt #275, Indianapolis, IN 46202  
Phone: (317)-909-2563, Email: [kchetal@iupui.edu](mailto:kchetal@iupui.edu)

## QUALIFICATION SUMMARY

- Proficient in web technologies and scripting languages. Build “Next Generation Sequencing pipeline” and “Statistical packages for R” which are used at research institutes for processing high throughput data
- Also worked on Genome Annotation for Human and Bacterial genomes and also implemented different computational methods for the analysis of these prokaryotic and eukaryotic genomes
- Extended knowledge of Windows, Linux/Unix and MAC OS X
- Developed various programs, some of which included file automation and differential expression using Python
- Developed different pipelines, tools for analysis of data and working with the data
- A highly enthusiastic & motivated team player with flexible thinking

## PROFESSIONAL EXPERIENCE

### Intern, Torrent Pharmaceuticals

June 2010

The training was based on molecular modeling and docking techniques. The docking study revealed the interacting domains of the Tax viral protein

### Intern, Bioinformatics Institute of India

June 2009

Training on Genomic Analysis of genes responsible for Cold Urticaria and In-silico analysis of disease inheritance, protein modeling and drug discovery

## RESEARCH EXPERIENCE

### Research Assistant, Janga Lab of Genomics and Systems Biology, IUPUI

Fall 2012 - Present

#### Thesis:

- Genomic Insights & RNA-sequencing analysis for Convergent Transcription in Bacterial Genomes. (Paper submitted for review)
- Developed a database for multi-gene operon for bacterial genomes from RNA-seq data. (<http://107.170.50.123/>)

## TECHNICAL PROJECTS

- Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants
- Analysis of Human and Yeast Transcription Factor Networks
- Pancreatic Cancer Research Database
- Evaluation of Various Motif Finding Tools
- Structural Inference of Shiga Toxin mutation in Shigella Dysenteriae (Undergraduate Thesis Project)

## **EDUCATIONAL QUALIFICATIONS**

- **Master's Degree in Bioinformatics** **September, 2014**  
Indiana University Purdue University Indianapolis, Indiana  
GPA: 3.50/4.00
- **Bachelor of Science, Bioinformatics** **May 2012**  
VIT University, Vellore, India  
GPA: 3.36/4.00

## **SKILLS**

- Programming Languages: Python, R, Shell Scripting, HTML, CSS, JavaScript
- Packages: Statistical Packages in R/Bioconductor for Microarray analysis
- Visualization Tools: Cytoscape, IGV, jBrowse, MeV, Java tree view
- NGS Analysis: SAM Tools, Rockhopper, Mass Tandem, MS-GFDB, MS-Align+, Tophat 2 & Cufflinks
- Operating Systems: Windows XP, Vista, 7/8, OSX, Linux

## **HONORS**

- Poster Presentation CTSI Indiana, Sept 2013 on "Genomic & Mechanistic Insights for Convergent Transcription in Bacterial Genomes"
- Poster Presentation Rustbelt RNA Meeting RRM, Oct 2013 on "Genomic & Mechanistic Insights for Convergent Transcription in Bacterial Genomes"
- Poster selected for GIW, South Korea, 2011 on "Modeling and Interaction Analysis of NDM1 protein to identify the amino acids involved in drug hydrolysis"
- Best Life Science Poster Award in Gravitas Tech Fest, VIT, Vellore, India
- Volunteered Pinnacle, Bioinformatics Technical Fest at VIT University, 2008
- Discipline committee coordinator at Riviera at VIT University, 2009