



Development of an Operon Detection Algorithm to Analyze Gene Regulation in Drug Resistant *Mycobacterium tuberculosis*

*A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in the South African National Bioinformatics Institute
(SANBI), University of the Western Cape.*

Tracey Calvert-Joshua

Student number: **3005444**

Supervisor: Professor Alan Christoffels

Date: 16 September 2022



KEYWORDS

Mycobacterium tuberculosis

Drug resistance

Rifampicin

RNA sequencing

Operon prediction

Condition Specific Mapping of Operons (COSMO)

Decision tree

Algorithm validation

efflux pumps

genotype-specific operons



UNIVERSITY *of the*
WESTERN CAPE

ABSTRACT

Development of an Operon Detection Algorithm to Analyze Gene Regulation in Drug Resistant *Mycobacterium tuberculosis*

T Calvert-Joshua

*PhD Bioinformatics Thesis, The South African National Bioinformatics Institute,
University of the Western Cape*

Background: In prokaryotes, operon structures often form to allow microorganisms to respond rapidly and efficiently to changing environmental conditions. Operons are sets of neighbouring genes which are co-regulated and co-transcribed. Studies have shown evidence of operons changing their lengths and/or maintaining their lengths while up- or downregulating their expression levels when exposed to various stresses. Since several operons have also been associated with drug resistance, having access to the operon map of *Mycobacterium tuberculosis* (*Mtb*), may give us insight into the existing mechanisms employed by *Mtb* to circumvent drug stress, and more importantly, it may allow us to target larger sections of a genome when designing antitubercular drugs. Although REmap was applied to the *Mtb* genome, none of the existing operon predictors, was optimized for the unique genome of *Mtb*. We therefore aimed to build a new operon predictor based on the foundation laid by REmap and extended the algorithmic parameters using empirical evidence. We also aimed to identify operons that were both modified in length and differentially expressed under rifampicin (RIF) stress and to observe if this was done in a genotype-specific or autonomous manner, by predicting operons for different *Mtb* genotypes.

Methods: We developed COSMO, an algorithm that uses features of the *Mtb* genome and RNA expression data. We verified four parameters by evaluating a set of 49 experimentally verified operons (EVOs) and a matching simulated operon set. Our expression, data-informed parameters were: i) a minimum coding sequence (CDS) coverage, ii) a minimum intergenic region (IGR) coverage, iii) a maximum fold difference (FD) between adjacent CDSs and iv) a maximum FD between an

IGR and its flanking CDSs. COSMO also has a built-in feature which evaluates the length of the operon upon the addition of each new CDS, by testing whether the averages of all CDS belonging to the operon are still within the FD cut-off.

Results: In verified operons, the coverages of IGRs were more upregulated than the untranslated regions (UTRs) ($p = 0.005$). However, they were on average, half the coverage of their flanking CDSs ($p = 0.001$). Taken together, this demonstrates that IGR coverage is a significant parameter, but that it should be independent of CDS coverage. FDs between adjacent CDSs were significantly lower in verified operons than in the simulated operons ($p = 0.0007$) - adhering to a maximum FD between 5x-7x. Similarly, the maximum FD between the IGRs and their flanking CDSs were generally below 5x in verified operons ($p = 0.04$ and $p = 0.005$, for plus and minus strand, respectively). We compared the predictions of COSMO for Beijing samples to two other operon predictors: REmap and Rockhopper. COSMO accurately predicted more full-length operons (60%) under control and experimental conditions than REmap (50%) and Rockhopper (48%). COSMO also predicted twice as many operons as DOOR 2.0. COSMO was also better at distinguishing operons predicted under control conditions from those predicted under RIF stress. When we combined lineage 2 and lineage 4 samples, the prediction rate increased to 70% of EVOs.

Our multiple linear regression analysis showed that one of our new parameters – maximum FD between IGRs and CDSs - had the greatest weighting on correct operon prediction and that the traditionally used maximum FD between adjacent CDSs was the least significant parameter.

We showed that in general *Mtb* tends to resist operon reorganization – even under RIF stress. Approximately 80% of operons had the same call under control conditions as under RIF stress. That is, within a specific genotype ($n = 40$) of strains had consensus calls for their operon lengths ($p = 1.4 \times 10^{-9}$). In the ~20% of cases when operon lengths were modified, the data showed that these strains were more likely to modify their operon lengths within their genotype than to do so strain-independently, under both control ($p = 0.0006$) and RIF stress conditions ($p = 0.01$). Similarly, except for the efflux pump, MmpS5/L5, most operons were not significantly differentially expressed under RIF stress. This pump was also shown

to be under selection pressure to remain the same length regardless of stress or genotype differences. A gene enrichment analysis showed that operons were often split at genes involved in lipid metabolism with the aim to slow down *Mtb*'s growth rate and prolong survival under stress. It also showed that *Mtb* preferred to extend operon lengths with regulatory proteins – more specifically with regulatory proteins which are associated with lipid biochemical pathways. ATP-related proteins were preferentially packaged into housekeeping operons and were under the most intense selection pressure. Finally, by using nine drug sensitive strains that were grown under hypoxia, COSMO was able to predict an operon that was never predicted under RIF stress but was confirmed to be associated with a hypoxia pathway.

Conclusion: COSMO has outperformed three of the best operon predictors in predicting full-length operons, in the accuracy of these predictions and in distinguishing operons under control conditions from those under experimental conditions.

16 September 2022



DECLARATION

I declare that *Development of an Operon Detection Algorithm to Analyze Gene Regulation in Drug Resistant Mycobacterium tuberculosis* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full name: Tracey Lynn Calvert-Joshua

Date: 16 September 2022

Signed *TCalvert-Joshua*



ACKNOWLEDGEMENTS

Smack in the middle of this PhD, I hit a point where this PhD punched me in the gut. My self-esteem crumbled. And up until today, I can't even really pin-point what led to it. I believe it was multiple factors which just culminated into what seemed like a "Moses and the Israelites in the desert" situation. After admitting this to Professor Alan Christoffels, he told me in not so many words, that I should view the situation like this. A Masters is about learning the tools which are available and about building new skills, while the PhD is about knowing that you have all these tools in your toolbox, and about innovatively combining what you know and have, to apply them to novel situations. It's not about knowing everything. I don't think he realizes how much comfort those words brought to me. Because as a now-transformed perfectionist, this was the start of my journey into adopting a growth mindset.

During this PhD I got the education that the old system never afforded me. I learnt that there isn't just one correct answer. That failing is necessary to pass. That learning isn't a competition. That it's important to network towards synergy by focusing on the talents of the collective. That I need to know my weaknesses, so that I can leverage the strengths of others. That the sentence of my life story still has a comma at the end. So, I will fill many more roles as I grow. That it's just as important to grow in character as it is to grow in my career.

My mindset will therefore be everything going forward. Will I believe in the old world of my old mindset? In a world that is unfriendly and resisting me, or will I choose to believe that everything is conspiring to bring me all that I desire? Will I focus only on what I still need to accomplish and forget to make time to sink deep into my blessings? No, I will start and end everything with gratitude. Not just by saying my thank yous, but by feeling it.

So, thank you Alan for letting me learn from you. Thank you to Peter van Heusden, SANBI's cyborg encyclopaedia, who lured me get involved in so many things, and then quietly stepped back. You may have way too much faith in me. Thank you to

SANBI's supporting staff Ferial Mullins and Maryam Salie whose doors were always open and with whom I had long chats about life and laughed with and complained to. Thank you for your warmth. To Fungiwe Xholi Mpithi who became my friend and my sister. You always smiled, and sometimes I feel that only I knew how much it took to give that smile. Thank you to my husband Larry. You sacrificed your time, our planned adventures, your spending money, because you gave it to me, etc., so that I could take advantage of what women were deprived of for years. You deserve a doctorate for being the embodiment of love. Thank you, mommy and daddy for raising me with all of your strength and wisdom. A special thank you for your genetic engineering (**). Thank you to my Source. I know that many chips had to fall into place for this equation to work.

Thank you to my funder, the South African Medical Research Council Bioinformatics Unit, who funded me via Prof Alan Christoffels (grant holder). There was no way I could have completed this PhD without your faithful contribution.

And as part of the growth mindset, I thank myself. Because even if everything had to be handed to me (which it wasn't), I could have chosen to squander it. Tracey, I will no longer judge you harshly. I will treat you kindly and let you explore the quantum field of possibilities. I believe you can do this thing. And the next. You've got this woman!

print("Hello **beautiful, new, friendly** world!")

Table of Contents

DEVELOPMENT OF AN OPERON DETECTION ALGORITHM TO ANALYZE GENE REGULATION IN DRUG RESISTANT <i>MYCOBACTERIUM TUBERCULOSIS</i>	I
KEYWORDS	II
ABSTRACT	III
ACKNOWLEDGEMENTS	VII
TABLE OF CONTENTS	IX
LIST OF FIGURES	XII
LIST OF TABLES	XIV
ABBREVIATIONS AND ACRONYMS	XV
CHAPTER 1	1
INTRODUCTION	1
1.1 WHY DO WE NEED TO PAY ATTENTION TO OPERONS IN <i>MYCOBACTERIUM TUBERCULOSIS</i> (MTB)?.....	1
1.1.2 <i>What are operons?</i>	1
1.2 TARGETING OPERONS	3
1.3 PROBLEM STATEMENT	4
1.4 AIM	4
1.4.1 <i>Main Objectives</i>	5
CHAPTER 2	6
LITERATURE REVIEW	6
2.1 THE <i>MYCOBACTERIUM TUBERCULOSIS</i> GENOME	6
2.2 CURRENT <i>MTB</i> STATISTICS.....	7
2.3 WHAT ARE THE IMPLICATIONS OF OPERONS IN DRUG RESISTANCE?	7
2.3.1 <i>Efflux pumps</i>	7
2.3.2 <i>Compensatory mutations</i>	11
2.3.3 <i>Other mechanisms of action of operons involved in drug resistance.</i>	12
2.4 MECHANISMS OF DRUG RESISTANCE	12
2.4.1 <i>The role of operons in Mtb pathogenesis and virulence</i>	13
2.5 EXPERIMENTAL IDENTIFICATION OF OPERONS.....	14
2.5.1 <i>Advantages of RNA sequencing</i>	16
2.6 WHAT ARE THE FEATURES USED IN EXISTING OPERON PREDICTION ALGORITHMS?	17
2.6.1 <i>Advancements of predictive approaches</i>	18
2.6.2 <i>Limitations of existing algorithms</i>	19

CHAPTER 3	22
CONDITION-SPECIFIC MAPPING OF OPERONS (COSMO) USING DYNAMIC AND STATIC GENOME DATA.....	22
ABSTRACT	22
1 INTRODUCTION	24
2 METHODS	26
A) STRAINS AND GROWTH CONDITIONS	26
B) RNA EXTRACTION AND SEQUENCING	26
C) TRIMMING AND ALIGNMENT	27
2.1 THE ALGORITHM DESIGN	27
2.1.1 <i>Defining and extracting genomic features</i>	<i>28</i>
2.1.2 <i>CDS, IGR and UTR coverages of real operons versus fake operons.....</i>	<i>29</i>
2.1.3 <i>CDS coverage cut-off.....</i>	<i>29</i>
2.1.4 <i>Fold difference between adjacent CDSs.....</i>	<i>30</i>
2.1.5 <i>Minimum IGR expression cut-off.....</i>	<i>30</i>
2.1.6 <i>Fold difference between IGR and adjacent CDSs</i>	<i>30</i>
2.1.7 <i>Intergenic distance</i>	<i>32</i>
2.1.8 <i>Motifs at the start or the end of an operon</i>	<i>32</i>
2.2 ALGORITHM VALIDATION.....	32
2.2.1 <i>Multiple Linear Regression Analysis</i>	<i>33</i>
2.2.2 <i>Comparison to existing algorithms.....</i>	<i>34</i>
2.3 DECISION TREE.....	35
3 RESULTS.....	38
3.1 DEFINING OPTIMAL PARAMETERS	38
3.1.1 <i>CDS, IGR and UTR coverages of real operons versus fake operons.....</i>	<i>38</i>
3.1.2 <i>CDS coverage cut-off.....</i>	<i>38</i>
3.1.3 <i>Fold difference between adjacent CDSs.....</i>	<i>39</i>
3.1.4 <i>Minimum IGR expression cutoff.....</i>	<i>40</i>
3.1.5 <i>Fold difference between IGR and adjacent CDSs</i>	<i>41</i>
3.1.6 <i>Intergenic distance.....</i>	<i>43</i>
3.1.7 <i>Motifs at the start or the end of an operon</i>	<i>44</i>
3.2 ALGORITHM VALIDATION.....	45
3.2.1 <i>Multiple Linear Regression Analysis</i>	<i>45</i>
3.2.2 <i>Comparison to existing algorithms</i>	<i>46</i>
4 DISCUSSION	56
5 LIMITATIONS AND FUTURE WORK	59
6 DATA AVAILABILITY STATEMENT.....	60

CHAPTER 4	61
GENOTYPE- AND CONDITION-SPECIFIC OPERON PREDICTION FOR MYCOBACTERIUM TUBERCULOSIS UNDER RIFAMPICIN STRESS	61
ABSTRACT	61
1 INTRODUCTION	63
2 METHODS	65
A) SAMPLE COLLECTION	66
B) RNA EXTRACTION AND SEQUENCING	67
C) TRIMMING AND ALIGNMENT	68
2.1 PREDICTING OPERONS.....	68
2.1.1 Total number of operons predicted.....	69
2.1.2 Overall static operons	71
2.1.3 Dynamic operons.....	72
2.2 DIFFERENTIAL EXPRESSION OF OPERONS	72
2.3 FUNCTIONAL ANNOTATION OF OPERON GENES	74
2.4 TESTING COSMO ON <i>MTB</i> STRAINS UNDER HYPOXIA.....	74
3 RESULTS.....	75
3.1 TOTAL NUMBER OF OPERONS	75
3.1.1 Variance of samples	75
3.2 STATIC OPERON CALLS.....	77
3.2.1 Functional Annotation (FA) of Static TPs	79
3.2.2 Functional annotation of Static False positives	83
3.2.3 Functional annotation of Static False Negatives.....	86
3.3 DYNAMIC OPERONS	87
3.3.1 Heterogeneous calls versus consensus calls for dynamic operons.....	88
3.3.2 Operon change genotype or strain-specific?.....	89
3.3.3 Functional annotation of dynamic operons	91
3.4 DIFFERENTIAL EXPRESSION OF OPERONS (DEO).....	95
3.5 TESTING COSMO ON <i>MTB</i> STRAINS UNDER HYPOXIA.....	96
4 DISCUSSION	97
4.1 LIMITATIONS OF THIS STUDY	100
5 FUTURE WORK.....	101
6 SUPPLEMENTARY MATERIAL	102
7 DATA AVAILABILITY STATEMENT.....	115
CHAPTER 5	116
CONCLUSION AND FUTURE WORK	116
REFERENCES	119

LIST OF FIGURES

Figure 1: Illustration of the structure of an operon. In prokaryotes, a block of genes located adjacent to each other are often regulated by a single promoter and is called an operon, which may be transcribed as a single polycistronic mRNA. While structural genes encode products that serve as cellular structures or enzymes, regulatory genes encode products that regulate gene expression. A repressor is a transcription factor which binds to the operator to inhibit the transcription of structural genes. Alternatively, activators enhance transcription by enabling RNA polymerase to bind to the promoter (Parker et al. 2016).2

Figure 2: Illustration of the central region of an intergenic region relative to flanking CDSs. Two CDSs may be separated by an intergenic region (IGR) such as depicted with the two adjacent CDSs: CDS 1 and CDS 2. CDSs may also overlap such as the case with CDS 2 and CDS 3. In the case where they do not overlap, we calculate the FD between an IGR and each flanking CDS. However, we do not consider the entire IGR – which is the entire black line extending from where CDS 1 ends, to where CDS 2 begins, but only the centre of the IGR (red block). The entire IGR is only used when the IGR is ≤ 4 bp.31

Figure 3: Flow diagram of COSMO's workflow. The algorithm takes a bam file, a GTF file, four user defined cutoffs, together with some coordinate information on the genome. It then adds a CDSs to an operon if it satisfies all four conditions; the average coverage must be: **a)** equal to or above the CDS cutoff, **b)** equal to or above IGR cutoff, **c)** less than or equal to the maximum FD between an IGR and its flanking CDSs, and **d)** less than or equal to the maximum FD between adjacent CDSs.37

Figure 4: Gene expression (coverage) of CDS and IGRs for real and fake operons (Plus strand). In this figure, UTR coverages are shown in green. CDS coverages are in blue and IGR coverages are in red. **A)** There was no relationship between the CDSs and the IGRs of fake operons. In fact, in many fake operons, neighbouring CDSs were not even transcribed. **B)** The general observation for real operons, was that the expression of UTRs started to pick up before and trail off directly after the operon was transcribed. There seemed to be a correlation in the expression levels between the CDSs and IGRs of real operons. The UTR expression levels of fake operons were no different to those of the real operons. MWU for plus and the minus strand: $p = 0.33$ and $p = 0.13$, respectively. **C)** The expression levels of this experimentally verified operon demonstrates that even in real operons, some genes can have low expression levels (below 5x). Hence a strict minimum cutoff may not be feasible. Allowing the user to define the cutoff is more suitable.39

Figure 5: Comparison of the FDs and average coverages of the genomic regions (CDSs, IGRs and UTRs) being analyzed in real versus fake operons. **A)** The average coverages of adjacent CDSs of real operons were compared to the CDSs of fake operons. For the fake operons, some extremely large data points were removed, for a better view of the box plot. The FDs of adjacent CDSs in real operons usually remained within 5x-7x of each other and were also determined to be statistically significantly lower than those of fake operons ($p = 0.0007$). In contrast, the FDs of adjacent CDSs of fake operons often exceeded 10x. **B)** The coverages of IGRs in fake operons showed no significant differences in expression levels compared to the UTRs of fake operons ($p = 0.2$). In contrast, the expression levels of IGRs in real operons were more upregulated than that of the UTRs ($p = 0.005$). **C)** The CDS coverages of real operons were on average double that of their intervening IGRs (plus: $p = 0.04$ and minus: $p = 0.005$). **D)** The FDs of the interquartile range (IQR), the total length of the IGR and the centre of the IGR were compared to those of their flanking CDSs. Although the FDs of both the total IGR length and the centre of the IGR generally remained below 5x, the centre was chosen as the parameter for IGR coverage, since it had far fewer outliers. Some outliers were removed for better visualization of the box plots.42

Figure 6: Comparison of Intergenic distance between real and fake operons.44

Figure 7: Comparison of operons predicted by COSMO, Rockhopper and REMap. **A)** The intersection of operons from COSMO, Rockhopper and REMap. Only one operon was uniquely predicted by REMap, while three operons were uniquely predicted by Rockhopper. Seven operons were predicted solely by COSMO. **B)** COSMO also predicted four operons as only expressed under control conditions, while five operons were predicted only under RIF stress. **C)** REMap was also able to distinguish two operons that were predicted under control conditions from the one specific to RIF stress. Rockhopper is not shown, because the algorithm only predicted combined differentially expressed operons.50

Figure 8: Outline of the study design for lineage 2 and 4 samples. A combined 64 samples were used in this study which spanned Lineage 2 and Lineage 4. At least three biological replicates were used for each of the four families, namely: WT, rpoB mutants, Family X and Beijing. Each Family X BR also had at least three TRs. A genotype was considered to be the family together with its MIC status. Therefore, while the WT strains was one genotype and the rpoB mutants were another, the Beijing and Family X families consisted of 2 genotypes each. Thus, there is a total of 6 genotypes in the figure above.....67

Figure 9: The 65 genes belonging to the static TP operons were submitted to STRING-DB for a functional annotation analysis. The red proteins were linked to transporting of molecules/protons/electrons involved in ATP-related processes (FDR 1.9×10^{-12}). The blue and yellow genes were involved in 'response to stimulus' and to 'pathogenesis' (FDRs: 0.00014 and 0.0003), respectively.80

Figure 10: STRING-DB PPI of true positive operons and their predicted false positive genes. **A)** All of the predicted FP proteins of the operon Rv1334 – Rv1336, except for Rv1338 (murl), form part of the molybdopterin cofactor (MoCo) metabolic process (FDR = 9.34×10^{-8}), including three of the proteins which belong to the EVO. The proteins belonging to the experimentally verified operon (EVO) also function in a novel cysteine biosynthesis pathway. However, the MoCo metabolic pathway has been shown to be closely associated with the cysteine pathway. **B)** The genes of the EVO, Rv3793 – Rv3795, and its four upstream FP genes, Rv3789 – Rv3792, have been shown to function in the same BP, namely cell wall organization (FDR = 1.60×10^{-9}). Three of the FP proteins (Rv3790 – Rv3792) have been shown to form part of the operon under two other drug stresses. These are the proteins shown in red. For the official gene names of these common gene names, see **Supplementary Table 2**.85

Figure 11: Functional annotation of the genes constituting the dynamic operons. The most common biological processes (BPs) to which the dynamic operon genes belonged were 'cell and cell wall processes', 'intermediary metabolism and respiration' and 'regulatory proteins', respectively. Proteins involved in **lipid metabolism** were the most frequent target sites where an operon would be split, leading to FNs. Operons were most often extended by **regulatory proteins**, leading to FPs.93

LIST OF TABLES

Table 1: Efflux pumps implicated in drug resistance.	9
Table 2: Performance of the three operon prediction tools: COSMO, REMap and Rockhopper.	48
Table 3: EVOs predicted by three algorithms.	52
Table 4: Average percentage of the Mtb's genome predicted as operons and single genes across the 12 samples.	55
Table 5: Three scenarios demonstrating the rules for assignment base on replicates.	69
Table 6: The mean, standard deviation and coefficient of variance of COSMO's predictions for samples of the same genotype.	76
Table 7: The correctly identified operons and the operons which were static across genotypes.	77
Table 8: The correctly identified operons and the operons which were static across genotypes (continued).	78
Table 9: The most common BPs for the 16 static TPs operons across the genotypes.	82
Table 10: The dynamic operons and their predictions under control conditions versus RIF stress, per genotype.	91
Table 11: The differentially expressed operons called by limma voom	95



UNIVERSITY of the
WESTERN CAPE

ABBREVIATIONS AND ACRONYMS

AI	artificial intelligence
AMPs	antimicrobial peptides
ARC	Agricultural Research Council
ATP	adenosine triphosphate
BP	biological process
BR	biological replicate
CDS	coding region
CGA	comparative genomics approach
COSMO	Condition Specific Mapping of Operons
DEGs	differentially expressed genes
DEO	differential expression analysis of operons
DR	drug resistance
EMB	ethambutol
ETH	ethionamide
EVO	experimentally verified operon
FD	fold difference
FDR	false discovery rate
FN	false negative
FP	false positive
FQ	fluoroquinolone
GO	gene ontology
HIV	human immunodeficiency virus
IGR	intergenic region
INH	isoniazid
IQR	interquartile range
L1	Lineage 1
L2	Lineage 2
L4	Lineage 4
L5	Lineage 5
L6	Lineage 6



L7	Lineage 7
MAE	mean absolute error
MDR	multi-drug resistance
MDR/RR-TB	multidrug- or rifampicin-resistant TB
MDR-TB	multi-drug resistant tuberculosis
MIC	minimum inhibitory concentration
MLR	multiple linear regression
MoCo	molybdopterin cofactor
MTBC	<i>Mycobacterium tuberculosis</i> complex
MWU	Mann-Whitney U
NGS	next generation sequencing
OFL	ofloxacin
PPI	protein-protein interactions
PPV	positive predictive value
PPV	positive predictive value
PSWMs	position-specific-weight-matrices
PZA	pyrazinamide
qRT-PCR	quantitative reverse transcription-polymerase chain reaction
RIF	rifampicin
R2	coefficient of determination
RMSE	Root Mean Square Error
RNA	ribonucleic acid
RNA-seq	RNA sequencing
ROC	Receiver Operating Characteristic
ROS	reactive oxygen stress
RPA	RNAse protection assays
rpoB	β subunit of RNA polymerase
SNPs	single nucleotide polymorphisms
SVM	support vector machine
TB	Tuberculosis
TFBS	transcription factor binding site
TMM	trimmed mean M-values

TN	true negative
TPs	true positives
TREAT	t-tests relative to a threshold
TSS	transcription start site
TUs	transcriptional units
UTR	untranslated region
WHO	World Health Organization
WT	wild type
XDR	extensively drug-resistant



CHAPTER 1

INTRODUCTION

1.1 Why do we need to pay attention to operons in *Mycobacterium tuberculosis* (*Mtb*)?

Bacteria may be the simplest free-living life-form, but their presence in nature is ubiquitous. However, the cost of surviving under often adverse conditions, necessitated the formation of molecular tools that are rare in other kingdoms of life. As with gene expression in higher organisms, bacteria do not simultaneously transcribe all their genes according to their genomic capability. Instead, the necessary genes or proteins are transcribed according to their need. This need can be dictated by both their internal and their external environment. (Seshasayee et al. 2009). In 1960, Jacob et al., (1960) discovered that bacteria can make this process more efficient by the formation of operons.

1.1.2 What are operons?

Operons are co-transcribed and co-regulated clusters of neighbouring genes. As shown in **Figure 1**, they often share a promoter, which allows them to be transcribed as a single polycistronic messenger RNA (mRNA) (Jacob et al. 1960; Jacob and Monod 1961). Although the genes constituting an operon are jointly transcribed, an operon may modify the number of genes included in that operon (Dam et al. 2007; Güell et al. 2009). Changes in their external environment are often used as cues to facilitate these operon length modifications. For example, Lee et al. (2014) showed that *Salmonella* uses a specific series of mechanisms to receive a signal of the distinct environmental stress from the host, and then responds accordingly by activating the appropriate virulence operon. Similar behaviour was observed in *Mycobacterium tuberculosis* (*Mtb*). Singh et al. (2003) showed that the *mymA* operon is induced in macrophages upon exposure to an acidic pH, which may allow

Mtb to persist in its host. Moreover, Bretl et al. (2012) revealed that the *Rv1813c-Rv1812c* operon in *Mtb*, is upregulated in response to hypoxia, nitric oxide, and carbon monoxide stresses.

These survival strategies are however, not just activated in response to the usual stresses inside the human host immediately after infection, but they may extend to new stresses brought on by antibiotics. A cloning study carried out by Silva et al. (2001) revealed that the P55 gene of *M. bovis* was identical to the Rv1410 gene of *Mtb*. This gene that encodes a membrane protein, is a part of the Rv1410-Rv1411 **operon**, that conferred both aminoglycoside and tetracycline resistance in mycobacteria.

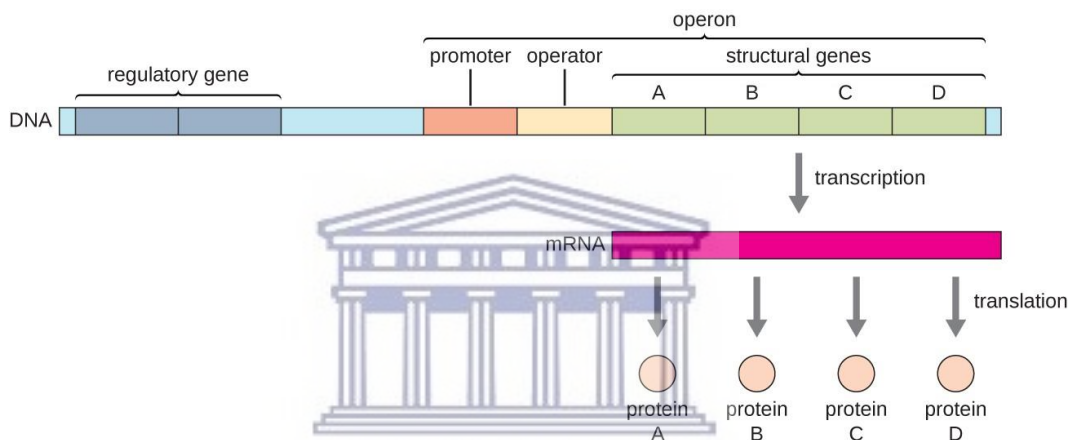


Figure 1: Illustration of the structure of an operon. In prokaryotes, a block of genes located adjacent to each other are often regulated by a single promoter and is called an operon, which may be transcribed as a single polycistronic mRNA. While structural genes encode products that serve as cellular structures or enzymes, regulatory genes encode products that regulate gene expression. A repressor is a transcription factor which binds to the operator to inhibit the transcription of structural genes. Alternatively, activators enhance transcription by enabling RNA polymerase to bind to the promoter (Parker et al. 2016).

1.2 Targeting operons

Numerous studies have shown the gross impact made on a phenotype by targeting entire *Mtb* operons (Thanassi et al. 1997; Banerjee et al. 1998; Pasca et al. 2004; Colangeli et al. 2005; Plinke et al. 2010). Shimono et al. (2003) showed that disruption of the entire *mce1* operon in TB-infected mice, resulted in a hypervirulent strain that was unable to enter a persistent state of infection – a state attributed to the success of this pathogen. Instead, the strains continued to replicate and kill their rodent host more rapidly than the wild-type (WT) strains. On the flip side, when two other operons, *mce3* and *mce4*, were completely deleted, this led to the attenuation of *Mtb* strains. The infected mice showed less prominent lung lesions at 15 weeks post-infection (Senaratne et al. 2008).

Experimental evidence therefore generously points to the significance of specifically targeting genes within an operon or entire operons to combat Tuberculosis. Continuing our research on a gene level, may not give us the advantage we need, because of compensatory mutations or the ability of the pathogen to switch to an alternative gene in the pathway.

However, compared to other pathogens such as *E. coli* and *Bacillus subtilis* (*B. subtilis*), the number of experimentally verified *Mtb* operons is small. Despite being an important human pathogen, the difficulties associated with targeting and culturing this slow-growing, fastidious species, have discouraged detailed genetic analysis. Recent non-culture-based molecular testing, such as RNA-seq, has the advantage of avoiding the delays of days to weeks for conventional culture (Lee et al. 1991; Yang and Rothman 2004; Speers 2006). Leveraging this data with bioinformatics approaches, removed another obstacle. The strength of combining molecular techniques with biological computation, has proven to be invaluable in rapidly advancing our insight into pathogenicity and virulence in prokaryotes. This recipe has since been applied to the identification of operons, leading to the prediction of several novel operons, which were later also experimentally confirmed (Roback et al. 2007; Pelly et al. 2016; Bundalovic-Torma et al. 2020; Tjaden 2020a).

1.3 Problem statement

Two of the predictors that were of special interest in predicting operons were REMap and Rockhopper (Pelly et al. 2016; Tjaden 2020a). REMap fared on par with DOOR 2.0 which outranked 14 other operon prediction algorithms. Rockhopper outperformed DOOR 2.0.

Both REMap and Rockhopper considered similarity in transcript abundance (coverage) between directly adjacent CDSs. REMap allows for a user to define an average cut-off for a CDS to be considered expressed, but the algorithm uses the same coverage cut-off for both the intergenic region (IGR) and the coding region (CDS). There is evidence that the coverages of these two genomic regions are not the same within operons. Rockhopper calculates the coverage of adjacent CDSs and their intervening IGRs but does not allow the user to define a cut-off. Due to insight into the organism's transcription and regulation, the user may want to define a minimum coverage. Rockhopper places a high significance on a short IGR distance, while REMap showed that this genomic region is inappropriate for *Mtb* operons. Furthermore, Rockhopper generates operons based on differential expression between a control and treated sample but does not show which operons are active under each condition. Lastly, neither predictor enforces a maximum fold change (FC) between adjacent CDSs or between adjacent CDSs and their intervening IGRs. Similarly, neither predictor uses a correlation of expression between all CDSs being added to a putative operon, as a feature.

1.4 Aim

In this thesis, we aimed to develop an algorithm to predict operons under control conditions versus RIF stress, using RNA-seq data. We developed our algorithm, Condition-Specific Mapping of Operons (COSMO), which is available at <https://github.com/SANBI-SA/COSMO>.

COSMO leverages the foundation laid by REMap and Rockhopper but offers additional parameters for improved operon prediction.

1.4.1 Main Objectives

Our main objectives of this study, were to determine:

- i) if there should be a minimum cut-off for the CDS and IGR to be considered expressed
- ii) if a separate expression level cut-off should be considered for the IGR and its flanking CDSs,
- iii) what the maximum fold differences between IGRs and their flanking CDSs should be,
- iv) whether we can improve operon prediction by considering the expression levels of all the CDSs of an operon and not just two immediately adjacent CDSs
- v) if COSMO was able to improve on the prediction rate of existing algorithms and
- vi) if COSMO could predict operons under RIF stress for different *Mtb* lineages

In Chapter 2 we discuss the uniqueness of the *Mtb* genome, implications of operons in drug resistance, the current methods for identifying operons experimentally and computationally, as well as why we need a new operon predictor for *Mtb*.

In Chapter 3 we show the steps involved in developing COSMO and we benchmark COSMO against REMap and Rockhopper.

Lastly, in Chapter 4 we predict operons for RIF treated and untreated strains from six different genotypes and analyze operon length changes and we evaluate differential expression of operons (DEO) under RIF stress. We also take a subsample of *Mtb* strains from a public database, which were grown under hypoxia stress, to ascertain if a unique set of operons can be predicted under a different environmental stress.

CHAPTER 2

LITERATURE REVIEW

2.1 The *Mycobacterium tuberculosis* genome

Within the family Mycobacteriaceae, *Mycobacterium* represent a genus of bacteria consisting of highly successful pathogens such as *Mycobacterium leprae* and *Mycobacterium ulcerans*, causing leprosy and Buruli ulcers, respectively. The *Mtb* species is one of the most common pathogens- causing tuberculosis (TB) in both humans and animals (Haning et al., 2014).

Mtb is a slow-growing tubercle bacillus that causes a chronic, infectious, airborne disease in susceptible patients. Its genome consists of a 4.4 megabase (Mb) circular chromosome constituting over 4000 genes in its guanine-cytosine-rich (G+C-rich) genome. This encodes 13 sigma factors, 11 two-component sensory transduction systems, 5 unpaired response regulators, 11 protein kinases and over 140 annotated transcriptional regulators. These transcriptional regulators have been implicated in response to stress signals and pathogenesis using mutagenesis and transcriptional profiling studies (Cole et al. 1998; Manganelli et al. 2004a; Arnvig and Young 2012). Its genome differs from those of other bacteria in that the greatest portion of its coding capacity is devoted to the production of lipogenesis and lipolysis enzymes and to two new families of glycine-rich proteins with a repetitive structure. This repetitive structure is believed to represent a source of antigenic variation (Cole et al. 1998). Antigenic variation is the mechanism used by infectious agents to evade the adaptive immunity of their host by altering their surface structures (van der Woude and Bäumler 2004; Coscolla et al. 2015).

Besides its slow growth, other characteristic features of this bacterium, is its dormancy, genetic homogeneity, intracellular pathogenicity and complex cell envelope. The generation time (time it takes to double in number) of the bacterium is ~24 hours, which contributes to its chronic disease nature, its lengthy treatment regimen and the challenges it imposes for researchers who wish to study it (Cole et al. 1998).

2.2 Current *Mtb* statistics

Until the COVID-19 pandemic, the *Mtb* pathogen was the leading cause of death amongst infectious diseases worldwide. Alarmingly, South Africa ranks as one of the ten countries that accounts for more than 90% of the global TB cases (World Health Organization 2022). The high burden of human immunodeficiency virus (HIV) infections further exacerbates the problem, as people living with HIV are on average 18 times more likely to develop active TB disease. In addition, the WHO 2019 statistics showed a 10% increase in multidrug- or rifampicin-resistant TB (MDR/RR-TB) cases, since the previous year (WHO 2020). This number increased again during 2021 (World Health Organization 2022).

2.3 What are the implications of operons in drug resistance?

2.3.1 Efflux pumps

The possibility of designing better anti-tubercular drugs against *Mtb* or of attenuating virulent *Mtb* strains, has garnered a special interest in the scientific community. One mechanism of action that seems to be prevalent in all prokaryotes, are efflux pumps. There has long been speculation that efflux pumps may be involved in both pathogenesis and virulence for the following reasons: i) they are pervasive in all living cells, ii) the genes encoding them belong to the bacterial core genome, iii) a single bacterial cell usually contains more than 10 different efflux pumps (redundancy), iii) they are non-specific; each efflux pump is able to export a variety of different substrates, and iv) their expression is tightly regulated (Webber et al. 2009).

Efflux pumps have been described in a wide variety of pathogens in which they generally exist as **operons**. Some of the most recognized efflux pumps and their resistance profiles are described in **Table 1**. One of the best studied efflux pumps in *Escherichia coli* (*E. coli*), the *AcrAB-TolC*, is a major contributor to intrinsic resistance to antibiotics in this micro-organism (Thanassi et al. 1997). Similarly, activation of this operon in *E. coli* is also the predominant cause of MDR in strains.

These transcriptional changes both downregulate cell influx and upregulate an intrinsic efflux system. Their mechanisms are hypothesized to be extendable to *Mtb*, since transformed cells expressed in *Mycobacterium smegmatis* displayed similar MDR profiles (Alekhshun and Levy 1997).

The *AcrAB-TolC* operon is a MDR efflux pump in *Salmonella enterica*. Its three genes contribute to additional virulence factors involved in the adhesion, invasion, and colonization of its host.

S. enterica mutants lacking these operon genes, showed differential expression of other major operons which were not only involved in virulence, but also in pathogenesis - supporting the hypothesis that there may be crosstalk between resistance and pathogenicity. Stated differently, operons involved in efflux systems, for which the functions may have been limited to causing disease (pathogenesis), may very well also be involved in virulence; especially virulence factors that play a role in antibiotic resistance.

Several other studies have also highlighted the contribution of efflux pumps in *Mtb* drug resistance. Pasca et al. (2004) showed that the Rv2686c-Rv2688c operon encode an ABC transporter responsible for fluoroquinolone efflux, while Colangeli et al. (2005) discovered that the *iniABC* operons are induced by isoniazid INH and ethambutol (EMB). A PCR experiment carried out by Plinke et al. (2010) showed that non-synonymous mutations in 15 distinct codons of the *embCAB* operon, which target all of the genes in this operon, were present in EMB-resistant strains.

Table 1: Efflux pumps implicated in drug resistance.

Organism	Drug resistance	Drug efflux operon	Reference(s)
<i>Acinetobacter baumannii</i>	aminoglycosides	<i>adeABC</i>	(Magnet et al. 2001)
<i>Bacillus subtilis</i>	thiolactomycin, microcin B17, sparfloxacin, carbonyl cyanide m-chlorophenylhydrazone (CCCP), tetrachlorosalicyl anilide, nalidixic acid	<i>emrRAB</i> and <i>mcbABCDEFG</i>	(Lomovskaya et al. 1995, 1996; Brooun et al. 1999; Xiong et al. 2000)
<i>Burkholderia pseudomallei</i>	aminoglycosides and macrolides	<i>amrAB-oprA</i>	(Moore et al. 1999)
<i>Escherichia coli</i>	hydrophobic antibiotics and detergents	<i>acrEF</i>	(Pan and Spratt 1994)
	Novobiocin and Deoxycholate	<i>mdtABC</i>	(Baranova and Nikaido 2002; Nagakubo et al. 2002)
	non-ionic detergent, hydrophobic agents (HAs), nonoxynol-9	<i>mtrCDE</i>	(Rouquette et al. 1999)

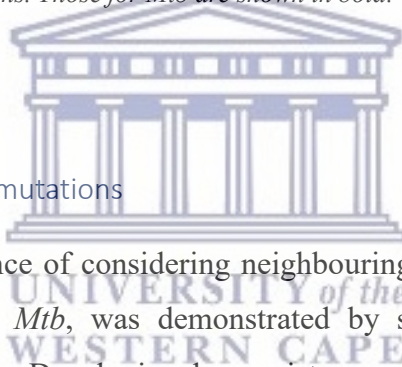
Table 1: Efflux pumps implicated in drug resistance (continued).

<i>Mycobacterium tuberculosis</i>	isoniazid (INH)	<i>furA-katG</i>	(Siu et al. 2014)
	INH	<i>mabA-inhA</i>	(Banerjee et al. 1998; Ando et al. 2014)
	ethambutol (EMB)	<i>embCAB</i>	(Plinke et al. 2010; Telenti and Iseman 2012)
<i>Neisseria gonorrhoeae</i>	hydrophobic agents	<i>mtrCDE</i> and <i>farAB</i>	(Lucas et al. 2014)
<i>Pseudomonas aeruginosa</i>	aminoglycosides, β -lactams, quinolones, chloramphenicol, tetracycline, trimethoprim, sulfamethoxazole, and novobiocin	<i>mexAB-oprM</i>	(Poole et al. 1996; Evans et al. 2001)
	quinolones, chloramphenicol and trimethoprim	<i>mexEF-oprN</i>	(Köhler et al. 1999)
	aminoglycosides	<i>mexXY</i>	(Aires et al. 1999)
<i>Pseudomonas putida</i>	tetracycline, chloramphenicol, carbenicillin, streptomycin, erythromycin and novobiocin	<i>arpABC</i>	(Kieboom and de Bont 2001)

Table 2: Efflux pumps implicated in drug resistance (continued).

<i>Staphylococcus aureus</i>	hydrophobic cations	<i>qacA/qacB</i>	(Grkovic et al. 2001; Schumacher et al. 2002)
<i>Stenotrophomonas maltophilia</i>	aminoglycosides, β -lactams, and fluoroquinolones	<i>smeABC</i>	(Li et al. 2002)
<i>Vibrio cholerae</i>	deoxycholate (DOC), chloramphenicol, nalidixic acid, and CCCP	<i>vceCAB</i>	(Woolley et al. 2005)

*An extensive number of efflux pump operons involved in antibiotic resistance have been discovered in many pathogens. Those for *Mtb* are shown in bold.



2.3.2 Compensatory mutations

Likewise, the importance of considering neighbouring genes when attempting to elucidate virulence in *Mtb*, was demonstrated by several studies focused on compensatory mutations. Developing drug resistance usually comes at a fitness cost to *Mtb* strains. For *rpoB* mutants, this means a reduced growth rate and a reduction in their overall virulence (Mariam et al. 2004; Rifat et al. 2017). However, several studies emphasized that mutations in the neighbouring *rpoA* and *rpoC* genes of *rpoB*, were shown to compensate for these *rpoB* mutations, restoring the fitness of strains by altering their gene expression in response to RIF (Comas et al. 2012; Naidoo and Pillay 2017; Xu et al. 2018).

2.3.3 Other mechanisms of action of operons involved in drug resistance.

Besides efflux pumps and compensatory mutations, other mechanisms of action, specifically activated by operons, have been linked to drug resistance. *Photothabdus laumondii* uses the *pbgPE* operon to resist being killed by antimicrobial peptides (AMPs) through **lipopolysaccharide modifications** (Derzelle et al. 2004; Bennett and Clarke 2005; Mouammime et al. 2017). Resistance to the frontline drug, EMB in *Mtb*, results from either mutations which cause the overexpression of the *Emb* protein(s), or from structural mutations in *EmmB*, or from both. This operon functions in the polymerization of **cell wall arabinan** (Telenti et al. 1997). As previously highlighted, Banerjee et al. (1998) showed that the upregulation of the *mabA-inhA* operon, led to an increase in the concentration of the *inhA* protein. This protein produces NADH-dependent enoyl-ACP reductase, which plays a role in **mycolic acid biosynthesis**. This series of events eventually produces the INH-ETH-resistance phenotype observed in *Mtb*.

2.4 Mechanisms of drug resistance

Despite research pointing to the targeting of operons in the early 2000s, subsequent research into understanding *Mtb* and designing antitubercular drugs, continued to focus on gathering information on single genes. For example, Chan et al. (2007) used nucleotide sequencing to investigate 213 multi-drug resistant tuberculosis (MDR-TB) clinical isolates, which were also resistant to two or more of the antitubercular agents: ofloxacin (OFL), rifampicin (RIF), ethambutol (EMB), isoniazid (INH) and pyrazinamide (PZA). MDR-TB is defined as being resistant to both INH and RIF (CDCTB 2016). Not surprisingly, they showed that a resistance phenotype was not always associated with the detection of a mutation in the corresponding known resistance gene. This was true even for the strains that were resistant to all five drugs, where mutations were regularly observed in only two or three of the many resistance genes. Interestingly, a proportion of clinical resistant isolates harboured no mutations in any of the drug resistance genes (Chan et al. 2007).

This was consistent with the findings of previous studies by Siddiqi et al. (2002; and Suresh et al. (2006). For example, RIF is the most important first-line antimicrobial used in the treatment of drug-sensitive tuberculosis and resistance to RIF is almost entirely due to mutations in its β subunit of RNA polymerase (*rpoB*). However, recent observations have implicated multiple other mechanisms for resistance to RIF. Many of these mechanisms appeared to be **downstream** of the initial trigger that promotes bacterial resistance (Zhu et al. 2018).

Similarly, in a more recent study, Bellerose et al. (2019) found that frameshift mutations in the *glpK* gene were found to be a specific marker of multidrug resistance (MDR) in clinical *Mtb* isolates. These loss-of-function alleles were not just limited to MDR but were also enriched in an extensively drug-resistant (XDR) clone. XDR can be defined as MDR-TB plus resistance to any fluoroquinolone and at least one of three injectable second-line drugs (i.e., amikacin, kanamycin, or capreomycin) (Cheon 2017). Similarly, genetic mutations in the single genes: *katG*-, *gyrA*- and *pncA*, have been associated with resistance to isoniazid, fluoroquinolone (FQ) and PZA, respectively. Yet again, these mutations were found to be present in many, but could not explain all resistance phenotypes (Suresh et al. 2006; Werngren et al. 2017; Zhu et al. 2018; Castro et al. 2020). Hence, it has become increasingly common to not be able to explain a drug resistance profiles with mutations in single genes.

2.4.1 The role of operons in *Mtb* pathogenesis and virulence

Meanwhile, in 2012, Hunt et al. demonstrated that *Mtb* needs the espACD-Rv3613c-Rv3612c operon for successful infection of its host. This highly antigenic operon must be precisely controlled because an incorrect level of its product, ESX-1, alerts the host's immune system. It has been suggested that variations in the expression of ESX-1, contributes to the diverse pathologies and their various host ranges. In fact, as far back as 1998, Banerjee et al. showed that the upregulation of

the *mabA-inhA* operon, led to the isoniazid-ethionamide (INH-ETH) resistance phenotype observed in *Mtb*. Yet, most current antibiotics are aimed at single genes. Unsurprisingly, the World Health Organization (WHO) has stressed its concerns over our failure to develop new antibiotics, stating that almost all new antibiotics brought to the market in recent decades, were just variations of those discovered in the early 1980s. Even those antibiotics currently in development, offer limited clinical benefit over the existing drugs. A rapid emergence of resistance to those drugs, which are not even on the market yet, is therefore to be expected. WHO has urged researchers to explore more innovative approaches to current antibiotic development, because failure to do so, may further fuel the impact of antimicrobial resistance (WHO 2021).

Targeting entire operon structures may be one such approach. This possibility was already demonstrated in a knockout study, where a mutation in the *mce* operon successfully attenuated virulent *Mtb* strains (Gioffré et al. 2005). Unlike genes, operons do not function in isolation, but tend to form part of higher-order biological modules (e.g., pathways). Mapping the operons in bacteria was therefore suggested to be essential for identifying novel pathways and biological processes, for assigning functions to hypothetical proteins and unknown genes that form part of an operon and for delineating operon promoters. The latter could lead to much larger and therefore more wide-ranging drug targets (Dandekar et al. 1998; Overbeek et al. 1999; Janga et al. 2005; Okuda et al. 2007; Bundalovic-Torma et al. 2020). These operon-related insights, may additionally give us profound understanding into the underlying mechanisms deployed by *Mtb* to persist, disseminate, cause disease, and develop drug resistance.

2.5 Experimental identification of operons

Currently, the most trusted methods to identify operons are by experimental studies. Although identifying operons experimentally is effective and generally precise, these types of studies are not as popular, partly due to factors that hinder their progress at a genomic scale. These factors include complexity, cost and duration (Walters et al. 2001). A review by Haller et al. (2010) reported that for 36 of the

experimentally verified operons (EVOs) in *Mtb*, several methods of operon identification were applied to gene pairs with longer operons; 16% of pairs were confirmed by primer extension, 15% of pairs used quantitative reverse transcription–polymerase chain reaction (qRT–PCR), 13% of pairs used promoter fusion experiments, 11% of pairs used Western or Southern blotting, and 15% of pairs used microarray co-expression (Ahmad et al. 2005; Roback et al. 2007; Casali et al. 2016). Some other common experimental methods to delineate operons include, but are not limited to, Northern blotting and RNase protection assays (Lynch et al. 2001; Sáenz-Lahoya et al. 2019).

Primer extension experiments attempt to locate the transcription start site (TSS) and/or potential promoters of an operon, using several specific primers that anneal to the 5' end of operon genes (Bagchi et al. 2005; Casart et al. 2008).

With qRT-PCR, cDNA from isolated RNA is reverse transcribed and the resulting product is viewed on an agarose gel for a size matching the mRNA (Woolley et al. 2005). In multiplex-PCR multiple target sequences are simultaneously co-amplified in a single reaction tube using more than one primer pair. The resulting amplicon may similarly be visualized by gel electrophoresis or be identified by hybridization with specific DNA probes and detected using spectrophotometry, fluorometry, autoradiography or chemiluminescence (Mahony and Chernesky 1995).

Promoter fusion experiments aim to analyze which genes/operons are induced by specific promoters. Promoter fusions are created by cloning a reporter gene in place of the TSS and promoter induction is measured by the expression of reporter genes (Hustmyer et al. 2018; Prezioso et al. 2018).

With Northern blotting, the total RNA of interest is resolved in a formaldehyde agarose and transferred to a nylon membrane. Blotted RNA is then separately probed (for annealed transcripts) with a radio- or non-radioactive labelled probe. The size of the product (band) is then visualized and measured by comparing it to a standard size ladder (Bhat et al. 2017). Southern and Western blotting follows a similar protocol to Northern blotting, but instead resolves DNA and protein, respectively (Singh et al. 2003; Bhatt et al. 2005).

RNAse protection assays (RPA) allows one to localize TSSs and to quantify mRNA expression levels. A single-stranded, discrete-sized, antisense probe is hybridized to an RNA sample. After hybridization, any remaining unhybridized probe and sample RNA are removed by digestion with a mixture of ribonucleases. Then, in a single step reaction, the nucleases are inactivated and the remaining probe:target hybrids are precipitated. These products are separated on a denaturing polyacrylamide gel and are visualized. RPAs are more sensitive than Northern blot analysis and are more accurate and direct than qRT-PCR analysis (Belin 1996; Lynch et al. 2001; Stacey et al. 2017).

2.5.1 Advantages of RNA sequencing

As previously stated though, experimental methods are costly and laborious at a genomic scale. However, those are just a few of the limiting factors. The nature of operons dictates that not all operons of an organism are formed and expressed simultaneously and across all environmental conditions (Dam et al. 2007). With laboratory experiments, only a few operons can be targeted at a time, which also means that researchers need to know which operons are induced at the time of the experiment. Microarray experiments aimed to solve this challenge by targeting multitudes of genes that are transcribed at a specific point in time and under certain conditions. RNA probes corresponding to gene sequences (oligonucleotides) are attached to a chip which are then used to capture the RNA present, by allowing the RNA to interact and bind to the probes. A microarray analysis also allows us to determine which genes are differentially expressed. However, even microarray experiments have the limitation that the genes of the organism must be known so that gene-specific probes may be created. Also, sufficient RNA must be present for detection and quantification of mRNA. As a result, even with microarray analysis, EVO data remained scarce (Parish et al. 2003; Wang et al. 2009; Mutz et al. 2013).

More recently, RNA sequencing (RNA-seq) has been considered the replacement for microarray expression studies to both map and quantify transcriptomes. RNA-seq is a rapid and inexpensive high throughput next generation sequencing (NGS)

technology. It offers the same ability as microarrays in providing us with a spatio-temporal snapshot of all genes expressed at the time of the nucleic acid extraction. However, compared to microarrays, RNA-seq has considerable advantages, including: the detection of novel transcripts and isoforms, measurement of allele-specific expression, and a large dynamic range of expression levels. It is not limited to designing and detecting only transcripts that correspond to a known genomic sequence. RNA-seq can also be used to resolve the exact locations of transcription boundaries up to a single-nucleotide resolution and to reveal sequence variations, such as single nucleotide polymorphisms (SNPs). Additionally, technical improvements have decreased sequencing costs which has drastically increased the size and number of available RNA-seq datasets (Wang et al. 2009; Mutz et al. 2013; Zhao et al. 2013).

Not surprisingly, this technology has already been used extensively in an attempt to understand *Mtb*'s host-pathogen interplay, the mechanisms behind XDR and the potential contributions of non-coding RNA in adaptive responses, amongst other insights (Arnvig et al. 2011; de Welzen et al. 2017; Pisu et al. 2020).

2.6 What are the features used in existing operon prediction algorithms?

Due to the explosion of RNA-seq data available in public archives, and the unfortunate cost and time associated with the experimental discovery of operons, several computational methods for operon predictions have been developed over the years. However, the current approaches for predicting operons vary immensely. Algorithmic features can roughly be divided into five categories: intergenic spacing, conserved gene clusters, functional relations, genome sequence based and experimental evidence (Jacob et al. 2005). Functional relations include gene functions, metabolic pathways, and protein-protein interactions. These types of operon predictors, rely heavily on a well-characterized organism and the accessibility of this data.

Genome sequence-based approaches include the use of codon adaptation indices, phylogenetic profiles, transcription start sites, promoters, terminators, transcription

factors and other motifs. Likewise, the use of these locations is complicated by the necessity for these positions to be well characterized - which is not the case for most microorganisms (Brendel and Trifonov 1984; Ozoline et al. 1997; Yada et al. 1999; Wang et al. 2007).

Finally, the experimental data used in operon predictors is usually comprised of microarray expression data. Although, more recently, RNA-seq data has become increasingly popular (Romero and Karp 2004; Price et al. 2005; Wang et al. 2007; Pelly et al. 2016; Tjaden 2020a).

Computational approaches are categorized into statistical analysis techniques, probabilistic scoring methods, rule-based prediction and artificial intelligence (AI) methods (Zaidi and Zhang 2017). Some probabilistic methods include, but are not limited to support vector machine (SVM), Bayesian, logistic regression, fuzzy scoring function, and neural network approaches (Bockhorst et al. 2003; Chen et al. 2004; Jacob et al. 2005; Price et al. 2005; Zhang et al. 2006).

2.6.1 Advancements of predictive approaches

At the start of the 20th century, Overbeek et al. (1999) and Ermolaeva et al. (2001) predicted operon **gene pairs** by conserved gene cluster analysis, using a comparative genomics approach (CGA). Meanwhile, Zheng et al., (2002) opted to exploit operon prediction by using metabolism-related genes. Although the recorded sensitivity (89%) and specificity (87%) of **gene pairs** were high, this method is highly dependent on biochemical pathway knowledge – which is not extensively available for most organisms. Also, many genes in confirmed operons are not functionally related, and so will be incorrectly excluded.

Bockhorst et al. (2003) chose the Bayesian network approach and was able to obtain a sensitivity and accuracy of 78% for predicting operon **gene pairs** in *E. coli*. This algorithm was improved with the addition of utilizing short intergenic distance as a feature, which further increased the accuracy to 88% and allowed for the prediction of 75% of the transcriptional units (TUs) of *E. coli* operons. It was also an important

study for shedding light on the significance of using intergenic distance in predicting operons. This feature would subsequently be used extensively in other operon predictors (Salgado et al. 2000). Finally, by minimizing the need for prior knowledge of the organisms genome, Laing et al., (2008) was able to achieve a high accuracy by implementing transcription factor binding site (TFBS) position-specific-weight-matrices (PSWMs) as operon delimiters, and achieved an accuracy of 83% in *E. coli* and 93% in *Streptomyces coelicolor*, by predicting their operon **gene pairs**.

Around the same time Zheng and Ermolaeva were working on their individual predictors (Ermolaeva et al. 2001; Zheng et al. 2002), Tjaden's group carried out a high-density oligonucleotide probe array analysis using *E. coli*, where they observed that using the correlated expression levels of two neighbouring genes (CDSs) was a reasonable indicator that those CDSs are co-transcribed (Tjaden et al. 2002). In addition, they found that using this correlated expression with the inclusion of similar expression levels of their intervening intergenic region, gave a much stronger signal for predicting operons. In 2013, McClure et al. worked with Tjaden to improve their Rockhopper algorithm by combining these features with a naïve Bayes classifier. They found that 90% of **gene-pairs** verified to be co-transcribed in RegulonDB, were predicted to be co-transcribed by their approach.

2.6.2 Limitations of existing algorithms

Despite these improvements, these methods all relied on predicting operon **gene pairs** or **TUs** and not whole operons. Predicting entire operons is more challenging and may therefore result in a large false negative rate. For example, even with the incorporation of multiple predictors such as: i) intergenic distance, ii) functional classification of genes to predict TU boundaries, iii) information on metabolic pathways, iii) protein complexes and iv) transporters, Romero and Karp (2004) was only able to make a moderate 4% improvement on their previous algorithm by correctly predicting 69% of operons. This was despite the use of a model organism

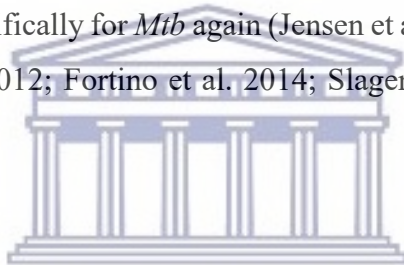
(*E. coli*) with the most extensive set of verified operons. Moreover, when this algorithm was tested on *B. subtilis*, the accuracy dropped by 23% (predicting 46% of verified operons), because of less available genomic information in the databases used.

A general problem with current methods is that they do not seem to generalize well from one genome to another. This may be due to a larger training or true positive set available for the one microbe. Alternatively, it may be a consequence of overfitting due to the use of genome-specific features (promoters, terminators, motifs, transcription factors etc.) that are unique to the one microorganism; leading to performance reduction, when applied to a new genome.

For example, in an earlier study Price et al. (2006) found that canonical spacing may not be under strong selection in *Escherichia coli*, *Salmonella* and some species of the *Bacillus* genus. In fact, adjacent CDSs in highly expressed verified operons tended to be widely spaced. In addition, the evolution of operons is also not always optimal, but more an adaptive approach. This is evident by more recently evolved operons which are comprised of functionally **unrelated** genes that were just in proximity before the operon was formed. A small IGR distance may therefore not be as indicative of co-regulation and co-transcription of adjacent genes as they anticipated. Similarly, Pelly et al. (2016) highlighted that the genome of *Mtb* is distinct from other prokaryotes in certain aspects of its genomic architecture. Unlike other bacteria, a great portion of its coding capacity is devoted to the production of lipogenesis and lipolysis enzymes and to two new families of glycine-rich proteins with a repetitive structure (van der Woude and Bäumlner 2004; Coscolla et al. 2015). *Mtb* also makes use of alternative sigma factors and show differences in -35 binding domains (Bashyam et al., 1996). In addition, 26% of *Mtb* genes produce leaderless transcripts, especially under **stress** (Cortes et al., 2013). As previously discussed, a short IGR is often the most important feature used in most operon predictors – including in that of Tjaden’s Rockhopper. However, even here, Pelly et al. (2016) confirmed that *Mtb* tends to have large IGRs between genes of verified operons. *Mtb* may therefore have an entirely different way of regulating its transcription.

Taking this into account, Pelly et al (2016) created REMap. To the best of our knowledge, REMap was the first to use RNA-seq data for *Mtb* operon prediction and to use only the IGR and CDS coverages as parameters. That is, REMap bypassed sophisticated machine learning methods which consists of training genomic data belonging to a specific organism. By using this simplified model, REMap was able to fair on par with DOOR, which was previously ranked as the best performing algorithm among 14 operon predicting algorithms (Mao et al. 2009). They were also able to predict strand-specific and condition-dependent operons and they predicted **full operons** – not just gene pairs.

Notably, between 2010 and 2020 several new operon predictors emerged using: differential RNA-seq data, functional relationships contained in STRING-DB and other databases, terminator sequences and statistical models; none of which predicted operons specifically for *Mtb* again (Jensen et al. 2009; Sharma et al. 2010; Taboada et al. 2010, 2012; Fortino et al. 2014; Slager et al. 2018; Taboada et al. 2018; Tjaden 2020a).



In summary, several studies have pointed to the significance of turning our attention to the role of operons in *Mtb*'s virulence – especially in the context of drug resistance. While experimental evidence is still currently the most trusted means of discovering operons, computational advancements have shown a great improvement in the accuracy of its operon predictions. Although several computational methods for operon predictions exist, the uniqueness of *Mtb*'s genome necessitates the development of an operon predictor that is not trained on the genomic architectures of other prokaryotes.

CHAPTER 3

CONDITION-SPECIFIC MAPPING OF OPERONS (COSMO) USING DYNAMIC AND STATIC GENOME DATA

Abstract

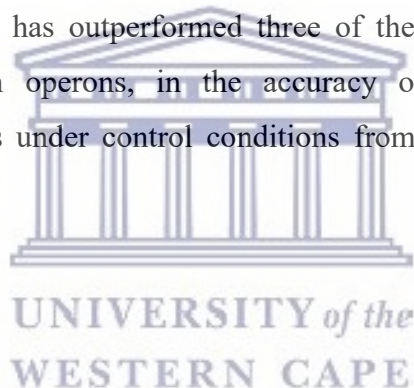
Background: An operon is a set of adjacent genes which are transcribed into a single messenger RNA (mRNA). These higher modules provide prokaryotes with a biological advantage to rapidly and efficiently circumvent both internal and external stresses. Approximately 60% of the *Mycobacterium tuberculosis* (*Mtb*) genome is believed to be arranged into operons. Having access to the operon network of *Mtb* may allow us to access larger sections of a genome for drug targeting. None of the existing operon predictors, except REMap, was created for the unique *Mtb* genome. We opted to improve on the foundation laid by REMap and Rockhopper and extended the genomic features with empirical evidence.

Methods and Results: We developed Condition-Specific Mapping of Operons (COSMO), an algorithm that uses features of the *Mtb* genome and gene expression data for *Mtb* exposed to RIF and its control. We verified four parameters by evaluating a set of 49 experimentally confirmed operons and a matching simulated operon set. Our first parameter was the minimum coding sequence (CDS) coverage, a parameter taken from REMap. The second parameter was the minimum intergenic region (IGR) coverage. In verified operons, the coverages of IGRs were more upregulated than that of the UTRs ($p = 0.005$). IGR coverage was also half the coverage of their flanking CDSs, demonstrating that IGR coverage is a significant parameter that should be used independently from CDS coverage. The third parameter was the maximum fold difference (FD) between adjacent CDSs. In real operons the maximum FD was between 5x-7x and was significantly lower than in fake operons ($p = 0.0007$). The maximum FD between and IGR and its flanking CDSs was the fourth parameter. In real operons, the maximum FD between IGRs

and adjacent CDSs were generally below 5x ($p = 0.04$ and $p = 0.005$, for plus and minus strand, respectively). Lastly, genes were also only added if they remained within the max FD of all genes already existing within a putative operon. A multiple linear regression analysis showed that our new parameter – maximum FD between IGRs and CDSs - had the greatest weighting on correct operon prediction and that the traditionally used maximum FD between adjacent CDSs was the least significant parameter.

We then compared COSMO to REmap, Rockhopper. COSMO accurately identified more operons under control and experimental conditions (60%) than REmap (50%) and Rockhopper (48%) and was also able to do so at a higher accuracy (75%), compared to REmap (67%) and Rockhopper (66%). We also compared COSMO to DOOR 2.0 and COSMO was able to predict twice as many operons as DOOR 2.0. Lastly, COSMO was also better at separating operons predicted under control conditions from those predicted under RIF stress.

Conclusion: COSMO has outperformed three of the best operon predictors in predicting full length operons, in the accuracy of the predictions and in distinguishing operons under control conditions from those under experimental conditions.



1 Introduction

Previous studies have shown that approximately 60% of the *Mycobacterium tuberculosis* (*Mtb*) genome may be arranged into operons (Pelly et al. 2016). An operon is a set of neighbouring genes that are co-transcribed as a single messenger ribonucleic acid (mRNA) (Price et al. 2006). These coregulated genes may not always be functionally related and may not always retain the same length (Osbourn and Field 2009). Under different conditions, an existing operon may be modified by the addition or the removal of one or several genes. Similarly, completely new operons may form, and old ones may be destroyed over time; demonstrating that operons are highly dynamic in their ability to evolve over both long and short periods of time. These changes often drastically alter the gene expression, and therefore also the phenotype of a species (Price et al. 2006; Güell et al. 2009).

However, if most genes in *Mtb* do not operate independently, then directing our anti-tubercular arsenal at individual genes may not be the most effective method of drug targeting. We need to be able to have an overview or network-level vantage-point to target larger segments of the genome more efficiently. In prokaryotes, operon structures often form in response to environmental stresses to aid these microbes to respond rapidly and efficiently in a bid to overcome adversity (Zaidi and Zhang 2017).

Fortunately, with the increasing availability of expression data in the form of RNA sequencing (RNA-seq), we are able to get an accurate representation of the overall gene expression within a species at any given moment (Zhao et al. 2013). RNA-seq is a more attractive approach than traditional platforms such as microarray, due to its wider dynamic range, its ability to predict more differentially expressed genes (DEGs), and to analyze a transcriptome at a single nucleotide resolution (Rao et al. 2019). With an overview of a microorganism's gene expression, we may be able to see which neighbouring genes are active and coregulated under orchestrated conditions. This data may allow us to predict operons for prokaryotes, rather than always resorting to costly and laborious experimental procedures.

Currently there are a plethora of operon predictors for prokaryotes. Tjaden (2020), who developed Rockhopper, one of the operon predictors against which our

algorithm was benchmarked, tested Rockhopper on ten commonly studied microorganisms. He demonstrated that even though experimental methods may be precise and provide strong evidence, many computational tools, such as Rockhopper, can now identify operon **gene pairs** with predictive accuracies that exceed 90%. Several operon predictors use genomic features such as pathway analysis, sequence homology, gene ontologies and intergenic region (IGR) distance (Che et al. 2006; Cao et al. 2019). Others steer away from being bound by existing genomic data and instead use statistical models to do their predictions (Bergman et al. 2007). However, to the best of our knowledge none of the existing operon predictors, except REMap, were optimized for the *Mtb* genome. The genome of *Mtb* has been shown to have some significant differences in its transcriptional preferences, such as the use of alternative sigma factors and differences in the -35 binding domains (Bashyam et al. 1996). In addition, 26% of genes produce leaderless transcripts. This was especially evident in strains under a stress model (Cortes et al. 2013). REMap also showed that longer IGR lengths are common between genes of *Mtb* operons, despite this being very uncommon in other prokaryotes. Short IGR lengths are often used as the most significant feature to identify operons in existing operon predictors (Salgado et al. 2000; Bergman et al. 2007; Chuang et al. 2012; Fortino et al. 2014; Taboada et al. 2018). These all indicate that *Mtb* has unique ways of regulating its transcription, which needs to be accounted for during algorithm design.

We have therefore used the foundation laid by REMap and Rockhopper, but improved upon a few areas of their design. We have developed an algorithm called ‘Condition Specific Mapping of Operons’ (COSMO), which uses our existing knowledge of operons and the structural annotation of the *Mtb* genome – which we call its static data. It also uses RNA-seq to get a snapshot of the gene expression profile observed when the *Mtb* is exposed- and not exposed to rifampicin (RIF) – which we call its dynamic data. These combined data sets are leveraged by COSMO to evaluate how operons may evolve in response to RIF stress.

2 Methods

a) Strains and growth conditions

Samples were obtained from two TB patients. The strains were classified according to their rifampicin minimum inhibitory concentration (MIC), with three having a high MIC (150 ug/ml) and three with a low MIC (40 ug/ml). The MIC is the lowest concentration of a substance, in this case an antibacterial, which results in either the maintenance or the reduction of a strain's growth (Lambert and Pearson 2000). They were classified as belonging to the Beijing genotype. Cultures were grown in 7H9 media until mid-log phase and exposed to RIF for 24 hours. The control batches received no RIF treatment. Both the high- and low MIC strains were exposed to a quarter MIC of RIF, resulting in a total of n = 12 samples. We chose 24 hours exposure and a quarter MIC_{RIF}, because the aim was to detect changes in transcription and not to kill the bacteria. The 24 hours also represents the doubling time of *Mtb* (Cole et al. 1998).

b) RNA extraction and sequencing

RNA extraction was carried out using the FastRNA Pro Blue kit (MP Biomedicals, Germany) and residual DNA was treated with DNase (Promega, WI, USA). Ribosomal depletion was performed with the bacterial option as probes for hybridisation of rRNA (TruSeq Total RNA, USA). Primer design and RNA-seq were carried out at the Agricultural Research Council (ARC) sequencing facility in Pretoria, South Africa, using the TruSeq DNA and RNA CD Indexes (I7 and I5 adapters) and the Illumina HiSeq 2500. The strand-specific protocol was confirmed as the fr-firststrand library type, using the *RSeQC v2.6.4* 'Infer Experiment' tool (Wang et al. 2012a).

The fastq files have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE203032 (samples GSM6152783 to GSM6152802).

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE203032>

c) Trimming and alignment

A Fastqc check in Galaxy (Jalili et al. 2020) showed that the data was of high quality (mean PHRED > 30), but the reads were nonetheless trimmed, and adapter sequences were removed using Trimmomatic V.0.38 with a PHRED = 20 with a sliding window of 4 (Bolger et al. 2014). Reads were aligned to H37Rv (NC_00096.3) using BWA-MEM 0.17.1 (Li and Durbin 2009). The quality of the bam files were checked using Samtools 1.9 (Li et al. 2009), to make sure ~90% or reads were paired, ~90% of reads were aligned to the reference genome, and that the average size of the reads were ~200bp, according to the RNA-seq analysis best practices (Conesa et al. 2016). Some of the bam files were converted to wiggle files for further analysis, using the *RSeQC* package in Galaxy (Wang et al. 2012b).

2.1 The Algorithm Design

We opted to improve on the foundation laid by REMap and Rockhopper by extending the algorithmic parameters with empirical evidence. With COSMO we verified **four parameters**, by evaluating a set of 49 experimentally confirmed operons and a matching simulated operon set – which we call the *fake* operons. The 49 operons were obtained from literature from both the plus (n = 30) and the minus strand (n = 19). We created a set of 49 fake/simulated operons, by using adjacent genes that were **not** previously confirmed to belong to an operon and which did not overlap with, and were not located in close proximity to verified operons. Each fake operon therefore matched its real operon counterpart by strand and by the number of genes. This true negative operon list was only used for this initial comparison between the coverages of real and simulated operons. We were always aware that they were not verified true negatives, but that they may have actual operons that have not yet been discovered. Therefore, this list was **not** used as a true negative (TN) set to measure the performance of the three algorithms.

Some of the parameters we aimed to include and assess for this algorithm were:

- i) how many reads should be available on average for a gene or CDS to be considered expressed (minimum CDS coverage)?
- ii) is there a correlation between expression levels of CDSs of the same operon? (maximum FD between adjacent CDSs)
- iii) Should the IGR be an independent parameter and if yes, then what should the minimum coverage be for us to consider it expressed (minimum IGR coverage)?
- iv) is there a correlation between the expression levels of an IGR and its flanking CDSs (maximum FD between IGR and its flanking CDSs)?
- v) should the entire IGR be used when we compare the IGR coverage to its adjacent CDSs or is a certain part of the IGR more tightly regulated with the CDSs?
- vi) should we use IGR length/distance as a feature?

The algorithm and all the scripts used in the testing and validation phases, can all be found at: <https://github.com/SANBI-SA/COSMO>.

2.1.1.1 Defining and extracting genomic features

The coverages were extracted for each CDS, IGR and UTR, for both the real and the fake operons, using the wiggle file. A wiggle file specifies the depth of aligned reads in a per-base format. The CDS is defined as the protein coding sequence, and for the purpose of this study, it was defined by the coordinates in the NC_00096.3 GTF file (Ensembl). The IGR was defined as the region between two adjacent CDSs on the same strand. The UTRs were taken as the regions 300 bases up- and downstream of the start- and end coordinates of operons. We used 300 bases, since this was below the maximum length of the longest IGRs in our dataset, and also the median value for long 3' and 5' UTRs (Arnvig et al. 2011; Sedlyarova et al. 2016). We exploited this data to compute the coverage depth, which is calculated using the number and length of reads mapped for each genomic position.

The Mann Whitney U (MWU) test was used to determine if there was a statistically significant difference in the average coverages and FDs, between the genomic regions of real and fake operons, using base R v3.6.1 (R Core Team 2021).

2.1.2 CDS, IGR and UTR coverages of real operons versus fake operons

First, we wanted to ascertain whether there were **observable differences** in the overall expression patterns of operons versus non-operons (fake operons). Based on the wiggle files generated from the aligned RNAseq reads, we drew plots in R Studio using base R v3.6.1 (R Core Team 2021). We hypothesized that in the real operons the coverages of adjacent CDSs and their intervening IGRs should be correlated, while they should show no correlation in the fake operons. The UTRs served as an additional control, because the UTR expression levels should technically not be under selective pressure for coregulation in both the real or fake operons, and therefore their expression levels were expected to **not** differ. In addition, by comparing the noncoding UTRs of real operons to the IGRs of that same operon, one should be able to observe that while the UTRs may be uncorrelated to the operon CDSs, the IGRs should show preferential correlation to their flanking CDSs. On the contrary, both the expression of the IGRs and UTRs should be uncorrelated to the adjacent CDSs in fake operons.

2.1.3 CDS coverage cut-off

The first genomic region which we considered, was whether there should be a minimum expression level cutoff for CDSs of **real operons (min-CDS)**, that could be set as a feature or parameter for operon prediction. If ***a*** is considered to be the start of a CDS, and ***b*** is considered to be the end of a CDS, then:

$$coverage = \sum_{i=1}^n r_i$$

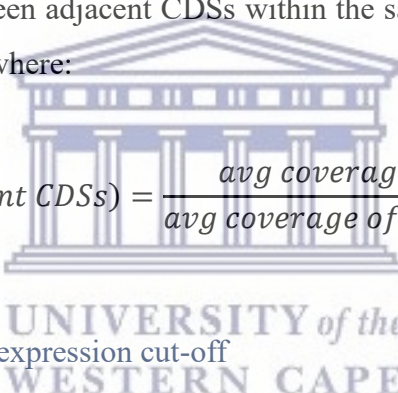
Where r_i is the total number of reads that mapped to the nucleotide position i , starting at position 1 in the genome, and n is the last position of the genome, or the length of the genome.

$$\text{Average coverage} = \frac{\text{coverage}}{(b - a) + 1}$$

The average coverages of the IGRs and UTRs were calculated similarly, using their specific genomic positions from the GTF file.

2.1.4 Fold difference between adjacent CDSs

To determine the correlation of expression levels between adjacent CDSs of operons, the FDs between adjacent CDSs within the same operon were calculated using a Python script, where:


$$FD(\text{adjacent CDSs}) = \frac{\text{avg coverage of a CDS}}{\text{avg coverage of adjacent CDS}}$$

2.1.5 Minimum IGR expression cut-off

We then considered the relevance of the IGRs. For the initial part of the analysis, we assessed if there was a statistically significant difference between the expression levels of IGRs versus UTRs in real operons compared to fake operons, since they are both non-coding regions. This was previously done only by observation. Then in the second part of the analysis, we tested whether we should use a minimum IGR coverage (**min-IGR**) as a user-defined parameter.

2.1.6 Fold difference between IGR and adjacent CDSs

We then assessed the relationship between the IGRs and their flanking CDSs. In computing the average expression level of an IGR, we need to bear in mind that expression levels increase before the transcription of a gene (i.e. before the 5' end)

and tail off after the 3' end. For IGRs longer than four base pairs we split the IGR into four regions and computed the average expression from the middle two segments, as shown in **Figure 2**, thereby avoiding the ramp up and tail off effects mentioned above. For IGRs shorter than four base pairs we computed the average expression from the entire IGR.

We recorded the FD for an IGR and its preceding CDS separate from the FD between an IGR and its succeeding CDS, to see if there was a difference. The FD between the IGR and its flanking CDSs was calculated as,

$$FD(IGR, CDS) = \frac{\text{avg coverage of IGR}}{\text{avg coverage of flanking CDS}}$$

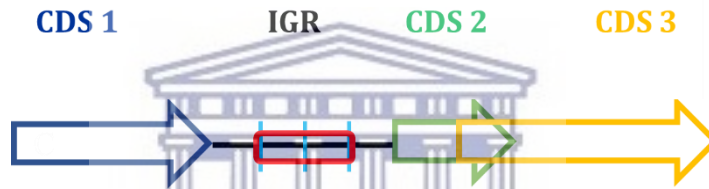


Figure 2: Illustration of the central region of an intergenic region relative to flanking CDSs. Two CDSs may be separated by an intergenic region (IGR) such as depicted with the two adjacent CDSs: CDS 1 and CDS 2. CDSs may also overlap such as the case with CDS 2 and CDS 3. In the case where they do not overlap, we calculate the FD between an IGR and each flanking CDS. However, we do not consider the entire IGR – which is the entire black line extending from where CDS 1 ends, to where CDS 2 begins, but only the centre of the IGR (red block). The entire IGR is only used when the IGR is ≤ 4 bp.

We recorded the FD for an IGR and its preceding CDS separate from the FD between and IGR and its succeeding CDS, to see if there was a difference. The FD between the IGR and its flanking CDSs was calculated as,

$$FD(IGR \rightarrow CDS) = \frac{\text{average coverage of IGR}}{\text{average coverage of each flanking CDS}}$$

2.1.7 Intergenic distance

We then compared the average IGR lengths in real operons to the average IGR lengths of fake operons. As previously stated, in many operon predictors, the intergenic distance is considered the most defining feature for accurate operon prediction. This is due to the finding that in most prokaryotes, the adjacent CDSs of operons often either overlap, or the IGR distances between adjacent CDSs are separated by fewer than 20bp of DNA. However, Pelly et al. (2016) reported that with *Mtb*, this distance was often 200 nucleotides long, but could reach up to 2.47kb in length. We calculated the lengths of IGRs of verified operons to observe if most IGRs were under 50bp. We used 50bp instead of 20bp to give some leeway to the longer IGRs. This was then used to consider it as a potential predictive feature.

2.1.8 Motifs at the start or the end of an operon

Finally, we hypothesized that if operons are under the control of a single regulator, then there may be a motif somewhere close to the operon to signal its start. Similarly, there may be a terminating sequence that signals the end of an operon. Since, in the case of promoters, this is usually upstream and downstream of a gene, we hypothesized that these signals may be up- or downstream of the operon. Hence, we extracted nucleotides up to 1000 bases up- and downstream of experimentally validated operons, using Pyfaidx 0.5.8. These nucleotides were submitted to Multiple Expectation maximizations for Motif Elicitation (MEME) to see if we could identify common motifs (Bailey et al. 2006).

2.2 Algorithm validation

Although previous cutoffs were statistically validated, we wanted to confirm these cutoffs by testing a range of **actual cutoff values**. We also wanted to ascertain if having all these options made a difference in terms of the total correct operon predictions. The algorithm was therefore run on nine Beijing lineage isolates, using permutations of the following cutoffs (python and bash scripts):

- a) min CDS cutoffs: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20]
- b) min IGR cutoffs: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20]
- c) max FD between adjacent CDSs: [5, 6, 7, 8, 9, 10, 15, 20]
- d) max FD between IGR and flanking CDSs: [5, 6, 7, 8, 9, 10, 15, 20]

This produced 9216 files per bam file/sample. A python script was used to compute the percentage of **full-length** operons from our EVOs list that were correctly predicted, or the proportion of true positives (TPs). We also calculated the percentage of false positives (FPs) and false negatives (FNs).

2.2.1 Multiple Linear Regression Analysis

We used the data from these 9216 files per sample, to perform a multiple linear regression analysis (MLRA) in R studio to observe whether each of the four parameters (independent variables), namely: minimum CDS, minimum IGR, FD of adjacent CDSs, and FD of the CDS compared to their IGR, had a statistically significant impact on the **outcome variable**. This was additionally used to verify **default** cutoff values. The outcome variable was the percentage of correctly predicted full-length operons, per combination of the different cutoff values. We performed a backwards stepwise analysis to remove any predictors that had no impact on the outcome. We also used the *'rpart'* (Therneau and Atkinson 2022) and *'rattle'* packages (Williams 2011) in R Studio to draw a pruned decision tree to confirm the default parameters in COSMO, for users who wish to run COSMO on default settings. Training and test sets were split into 70% and 30%, respectively. We reported the significant predictors, as per their t-statistic p-values, R^2 , the mean absolute error (MAE) and the Root Mean Square Error (RMSE). R^2 is the proportion of variation in the outcome variable that can be explained by the independent variables. The mean absolute error is an error statistic that averages the distances between each pair of actual versus observed data points (residuals) (Boiroju 2011). The RMSE gives us the standard deviations of the residuals from a model. This is often argued to be the more meaningful measure of a model's fit than

the R^2 metric (Alexander et al. 2015). The usual practice is to choose the model which has a lower accuracy measure among alternative models (Boiroju 2011).

2.2.2 Comparison to existing algorithms

Finally, the performance of COSMO was tested against other existing algorithms that use RNA-seq data as input. Initially we wanted to use DOOR 2.0, because DOOR was previously ranked as the best performing algorithm among 14 operon predicting algorithms (Mao et al., 2009). Unfortunately, it had become obsolete at the time of our testing. We therefore compared our results to Rockhopper and REMap. REMap was chosen because the algorithm's approach was similar to ours and tailor-made for the *Mtb* genome. Further encouragement was also due to REMap's performance which fared on par with DOOR 2.0 (Pelly et al. 2016). Rockhopper was our second comparator, because it was previously shown to outperform DOOR 2.0 (Tjaden 2019).

REMap published that an expression level of 10x was able to yield the best results. The algorithm however had a default value of 20x. Therefore, both parameters were used to predict operons using our datasets. Rockhopper does not allow user-defined expression level cutoffs.

The total number of TPs, as well as the total number of FPs and FNs, were calculated for each predictor. The algorithms were evaluated with the performance metrics: precision/positive predictive value (PPV), recall (sensitivity), and F1 score, where:

$$Precision/PPV (\%) = \left(\frac{TP}{TP + FP} \right) * 100$$

$$Recall/Sensitivity (\%) = \left(\frac{TP}{TP + FN} \right) * 100$$

$$F1 \text{ score} = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right)$$

In the absence of a verified true negative operon set, we could not calculate the specificities, accuracy scores or ROC curves. Secondly, the actual number of operons compared to single genes most likely results in an unbalanced dataset. In these two scenarios, the F1 score has been proven to be a better metric than the accuracy score to evaluate algorithm performance. Similarly, when datasets are unbalanced, precision and recall were demonstrated as better evaluators than sensitivity and specificity for a model's classification performance; and precision-recall curves were more useful and robust than the ROC curves (DeVries et al. 2021).

2.3 Decision Tree

After the selected features and parameters were statistically validated, COSMO was designed using the decision-tree based classifier. The decision tree classification method was previously tested in many different operon predictors and found to produce the highest sensitivity and specificity values (Chuang et al. 2012). It takes in a BAM file, a GTF file, as well as four user-defined parameters: *a*) a minimum CDS cutoff, *b*) a minimum IGR cutoff, *c*) the maximum FD between an IGR and its flanking CDSs and *d*) the maximum FD between two adjacent CDSs.

As shown in **Figure 3**, COSMO starts by checking if the first CDS it encounters is expressed. That is, it checks if the average CDS coverage is equal to or above the user defined CDS cutoff (*a*). If it is expressed, it then assigns it as CDS 1 of a putative operon. It then advances to CDS 2 on the same strand. If CDS 2 is expressed and it overlaps with CDS 1, it automatically gets added as CDS 2 of the operon. Should CDS 2 not overlap CDS 1, then the IGR has to be expressed. That is, the average coverage of IGR must be equal to or above the user-defined cutoff (*b*). The average of the entire IGR region is considered if its length is below 4 nucleotides. In the case where the IGR between CDS 1 and CDS 2 exceeds 4 bases, the IGR is split into four parts and the average of the bases making up the two middle regions is used for further analysis. Next, the FD between this IGR coverage and each adjacent CDS must not exceed the maximum IGR to CDSs cutoff (*c*).

Finally, before adding CDS 2 to the operon, the FDs between CDS 1 and CDS 2 must not exceed the maximum CDS-to-CDS cutoff (*d*).

The same process is followed with CDS 3. However, from CDS 3 and onwards, an additional rule is applied. For this fifth variable, the maximum FD is not just checked between CDS 3 and CDS 2, but the FD between CDS 3 and CDS1 must also not exceed the max-FD. If it does, then the operon is paused. However, COSMO does not automatically assume that because the coverages of CDS 1 and CDS 2 previously correlated, that they should remain an operon, and that CDS 2 should be the start of a putative new operon. Since the correlation ended when CDS 3 was compared to CDS 1, COSMO will conclude that the problem lies at CDS 1. It will therefore make a decision about how the operon should be split at CDS 1. It will evaluate whether the coverage of CDS 1 correlates better with CDS 2 and should therefore result in a bicistronic operon CDS 1 + CDS 2 or whether CDS 2 correlates better with CDS 3 and therefore become the bicistronic operon CDS 2 + CDS 3, with CDS 1 being expressed independently. This variable was a built-in feature in COSMO which could not be tested in the MLR. However, we suspected that this is likely to have a significant effect on the outcome variable.

Lastly, COSMO accounts for the circular chromosome of *Mtb*, so the first and last CDSs of the genome, may also form an operon. COSMO predicts strand specific, condition-dependent operons and outputs a CSV file. The output file contains the operon name and coordinates, the operon length and the average coverage of the operon, as well as the name and coverages of each individual gene and IGR within the operon.

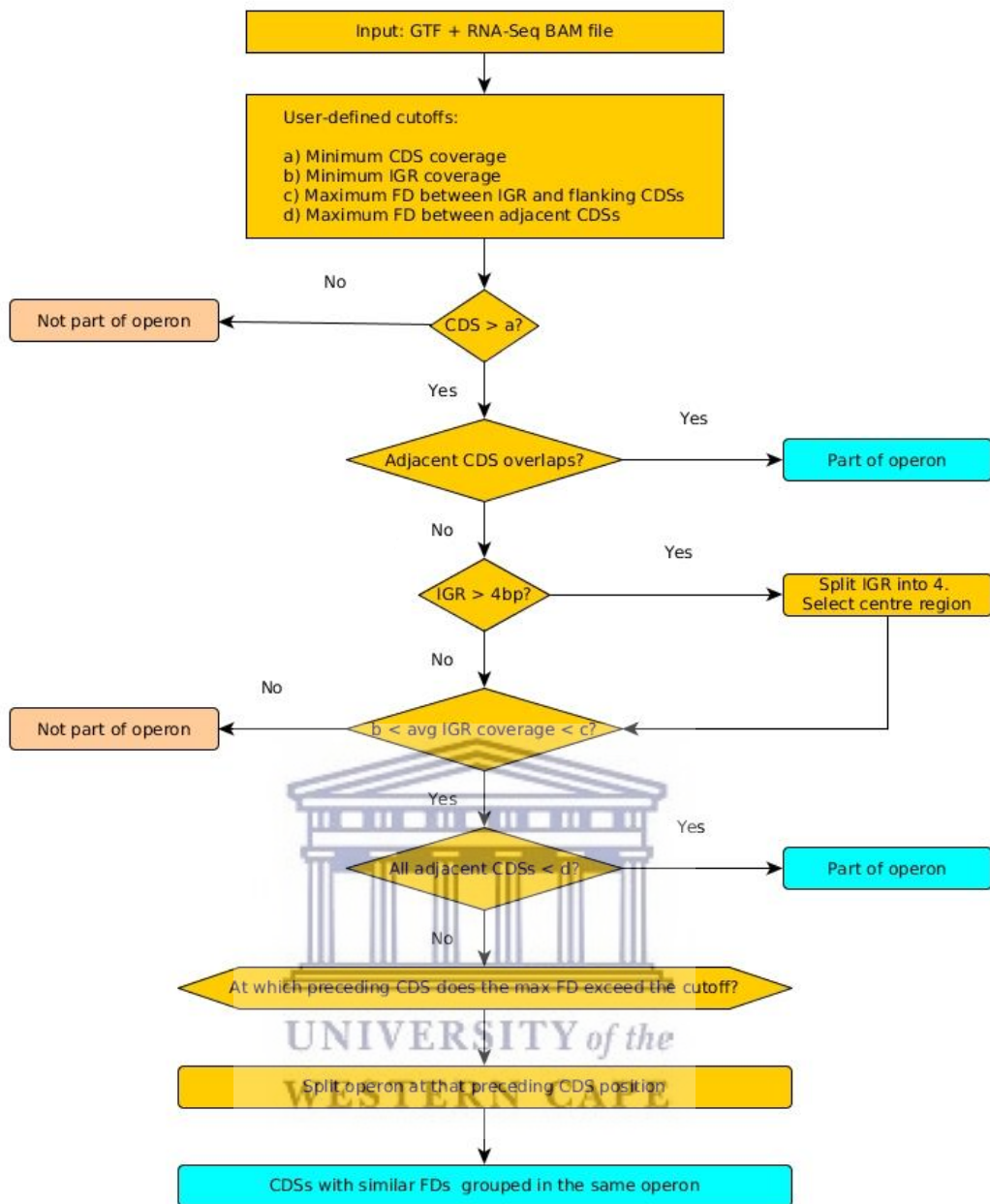


Figure 3: Flow diagram of COSMO's workflow. The algorithm takes a bam file, a GTF file, four user defined cutoffs, together with some coordinate information on the genome. It then adds a CDSs to an operon if it satisfies all four conditions; the average coverage must be: **a)** equal to or above the CDS cutoff, **b)** equal to or above IGR cutoff, **c)** less than or equal to the maximum FD between an IGR and its flanking CDSs, and **d)** less than or equal to the maximum FD between adjacent CDSs.

3 Results

3.1 Defining optimal parameters

3.1.1 CDS, IGR and UTR coverages of real operons versus fake operons

We used the wiggle file to plot the **raw** coverages of the individual bases, to observe whether there was consistency in the expression patterns across genomic regions, for the 49 real operons, which did not exist in the 49 fake operons. Unsurprisingly, as displayed in **Figure 4A**, the coding sequences (CDSs) (**blue**) and their adjacent IGR coverages (**red**), showed no correlation in *fake* operons. In contrast, **Figure 4B** shows that the CDSs of *real* operons generally showed a correlation in expression levels with their adjacent CDSs, as well as with their intervening IGRs. The untranslated regions' (UTRs) expression levels were no different between real and fake operons or single genes (plus and the minus strand: $p = 0.33$ and $p = 0.13$ respectively). The UTRs were later compared to the IGRs in **Section 3.1.4**. The UTRs served as controls to show that although the IGRs are also noncoding regions like the UTRs, within real operons IGRs are preferentially regulated and the UTRs are not, and therefore significant.

3.1.2 CDS coverage cut-off

The difference in expression levels between the CDSs of real versus fake operons were not statistically significant, for both the plus- and minus-strand (Mann-Whitney U test [MWU]: $p = 0.22$ and $p = 0.65$, respectively). This suggests that CDSs that make up operons are not necessarily targeted for upregulation, any more than independent CDSs (or single genes). Some of the CDSs of real operons were also expressed at very low levels - some were even below 5x coverage, as shown **Figure 4C**. Therefore, setting a high static CDS cutoff as a predictive feature, could cause the algorithm to bypass lowly-expressed or deliberately downregulated operons. The better solution might be to determine if there was a correlation of expression between adjacent CDSs of the real operons that does not exist within fake operons.

A fixed minimum value for the CDS coverage was therefore excluded as a static feature, but rather implemented as the *first user-defined parameter* in the algorithm. This is discussed later in **Section 3.2**

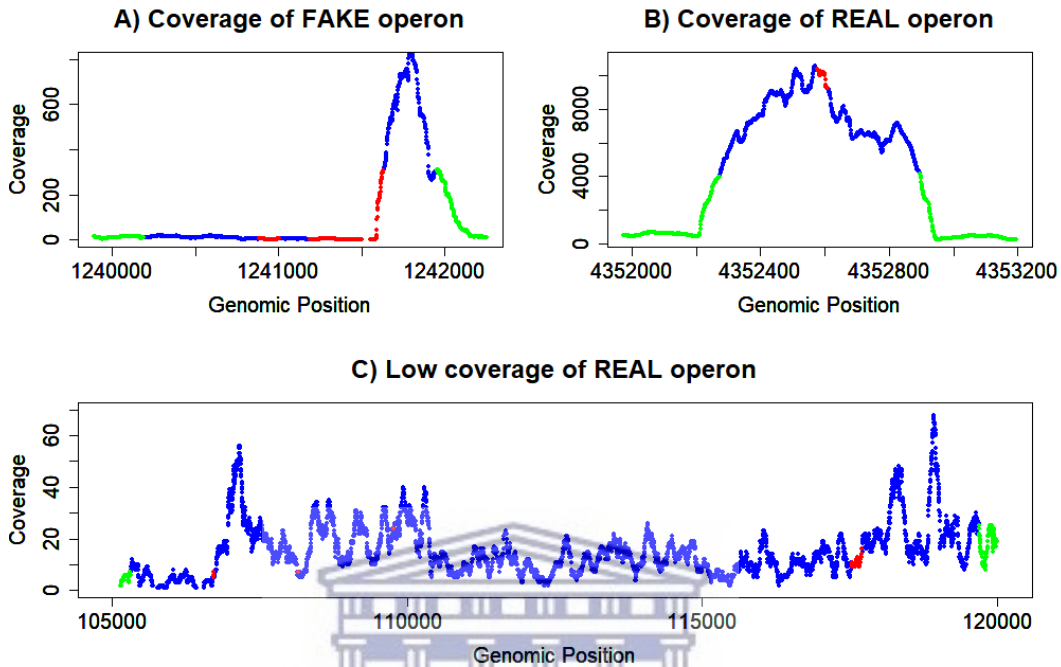


Figure 4: Gene expression (coverage) of CDS and IGRs for real and fake operons (Plus strand). In this figure, UTR coverages are shown in green. CDS coverages are in blue and IGR coverages are in red. **A)** There was no relationship between the CDSs and the IGRs of fake operons. In fact, in many fake operons, neighbouring CDSs were not even transcribed. **B)** The general observation for real operons, was that the expression of UTRs started to pick up before and trail off directly after the operon was transcribed. There seemed to be a correlation in the expression levels between the CDSs and IGRs of real operons. The UTR expression levels of fake operons were no different to those of the real operons. MWU for plus and the minus strand: $p = 0.33$ and $p = 0.13$, respectively. **C)** The expression levels of this experimentally verified operon demonstrates that even in real operons, some genes can have low expression levels (below 5x). Hence a strict minimum cutoff may not be feasible. Allowing the user to define the cutoff is more suitable.

3.1.3 Fold difference between adjacent CDSs

As anticipated, the fold differences (FD) of adjacent CDSs were more tightly regulated with respect to each other (**FD CDSs**) when they formed part of an operon, as displayed in **Figure 5A**. The maximum FDs between CDSs of real operons, were generally lower than those of fake operons ($p = 0.0007$). Adjacent

CDSs usually adhered to a maximum FD of 5x-7x. This threshold existed not just for a CDS and its immediately adjacent CDSs, but between all the CDSs that constituted that operon - even if the operon was up to 14 CDSs long.

In contrast, the FDs of adjacent CDSs within fake operons had a larger spread, many of which also frequently exceeded 10x, or even surpassed 20x (some outliers were removed). There was however one outlier for the real operons. The FD between genes Rv3418c and Rv3419c, of the operon Rv3417c-Rv3423c, exceeded 30x. However, previous literature demonstrated that under certain stresses, this operon could be split into two operons, namely: Rv3417c-Rv3418c and Rv3419c-Rv3423c. Thus, operon Rv3417c-Rv3418c, also known as groEL1-groES, is often expressed as an independent bi-cistronic operon, with the CDS Rv3418c showing evidence of gross upregulation in two experimental studies (Stewart et al. 2002; Aravindhan et al. 2009; Bhat et al. 2017). This was in alignment with our analysis. As a result of this exception in our already small test set, and because some operons may be better predicted with a slightly lower (or even a higher FD), we decided that we would also not restrict this value to a static maximum cutoff. Therefore, as the *second parameter* of the algorithm, users may choose their own maximum FD cutoff for adjacent CDSs, although we do advise to keep this value to a maximum of 7x. A default FD of 5x was built into COSMO if the user does not provide their own cutoff. The excel sheet and the graphs for all the operons, can be found on our GitHub page at:

https://github.com/SANBL-SA/COSMO/blob/master/Supplementary_data/ave_CDS_IGRs_for_COSMO_creation.xlsx

3.1.4 Minimum IGR expression cutoff

Regarding the IGRs, **Figure 5B** shows that in real operons, the coverages of IGRs were more *upregulated* than that of the UTRs ($p = 0.005$). As expected, when we similarly compared the coverages of the IGRs and UTRs for the fake operons, there was no statistical significance in their expression levels ($p = 0.2$).

This is in accord with our previous observations, which showed that while the CDSs and IGRs of real operons were tightly regulated - possibly by the same regulator - the UTRs are not. Additionally, the IGR coverages of real operons were also not the same as the CDSs but were on average 50% lower than the coverages of their flanking CDSs, as depicted in **Figure 5C** ($p = 0.04$ and $p = 0.005$; plus- and minus-strand, respectively). Therefore, the IGR coverage is not just a significant parameter when contrasted with the UTRs, but it should be an independent parameter relative to their CDSs. IGR coverage was therefore included as the *third user-defined parameter* in COSMO.

We then wanted to obtain a minimum cutoff for an IGR to be considered expressed. However, just as with the CDSs, the IGRs of real operons were generally not more up- or downregulated compared to individual IGRs of fake operons. The MWU test showed that the outcome was inconclusive. There was a statistically significant difference for the minus strand ($p = 0.01$), but not for the plus strand ($p = 0.14$). However, even though there may not be a defined minimum cutoff for the IGR coverage, from **Figure 4B** in the previous Section 3.1.2, we saw that in real operons, the IGRs (red line) show a correlation of expression levels with their adjacent CDSs. In contrast, we also previously showed in **Figure 4A** of Section 3.1.2, that the expression levels of IGRs and adjacent CDSs, behave haphazardly in fake operons. This suggests that just as with the CDSs of real operons, the IGRs may stay within a maximum FD to their adjacent CDSs.

3.1.5 Fold difference between IGR and adjacent CDSs

As depicted in **Figure 5D**, the possibility of using FDs between the interquartile ranges (IQRs) of the IGRs and their flanking CDSs were immediately discarded, because the spread of the data points representing the FDs, was too large and too random. Next, the FDs for the total IGR length and for that of the centre of the IGR were analysed. The boxplot shows that either one of the two may have been used as a source to calculate the FDs between the IGR and its flanking CDSs, because they did not perform very differently ($p = 0.49$). However, when the total IGR lengths were used, there were more outliers. Hence, the algorithm utilizes the centre of the IGR to establish the FD between an IGR and its adjacent CDSs. The

maximum FD for an IGR and its flanking CDSs was therefore also included as the *fourth user-defined parameter* of the algorithm.

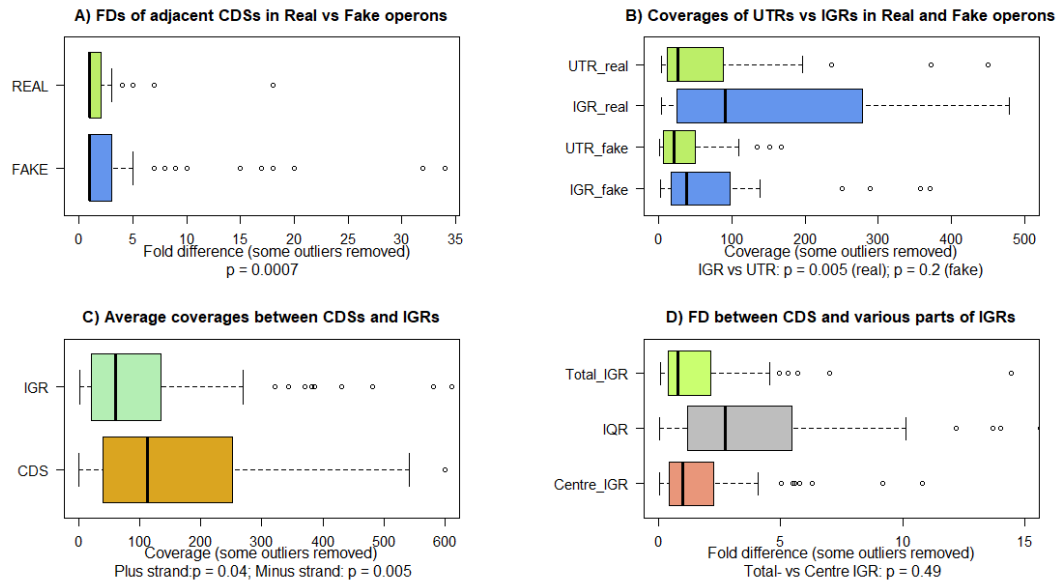
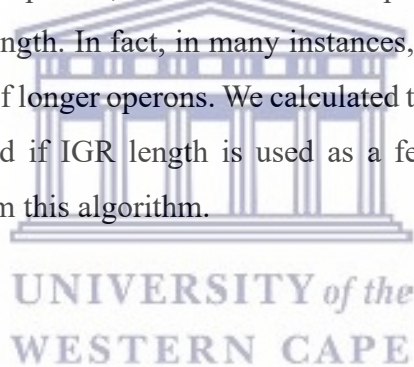


Figure 5: Comparison of the FDs and average coverages of the genomic regions (CDSs, IGRs and UTRs) being analyzed in real versus fake operons. A) The average coverages of adjacent CDSs of real operons were compared to the CDSs of fake operons. For the fake operons, some extremely large data points were removed, for a better view of the box plot. The FDs of adjacent CDSs in real operons usually remained within 5x-7x of each other and were also determined to be statistically significantly lower than those of fake operons ($p = 0.0007$). In contrast, the FDs of adjacent CDSs of fake operons often exceeded 10x. **B)** The coverages of IGRs in fake operons showed no significant differences in expression levels compared to the UTRs of fake operons ($p = 0.2$). In contrast, the expression levels of IGRs in real operons were more upregulated than that of the UTRs ($p = 0.005$). **C)** The CDS coverages of real operons were on average double that of their intervening IGRs (plus: $p = 0.04$ and minus: $p = 0.005$). **D)** The FDs of the interquartile range (IQR), the total length of the IGR and the centre of the IGR were compared to those of their flanking CDSs. Although the FDs of both the total IGR length and the centre of the IGR generally remained below 5x, the centre was chosen as the parameter for IGR coverage, since it had far fewer outliers. Some outliers were removed for better visualization of the box plots.

3.1.6 Intergenic distance

We then evaluated if IGR distance is an appropriate feature/parameter for *Mtb* operon prediction. As depicted in the density plot in **Figure 6A**, the peaks confirmed that most CDSs overlapped in real operons (had no IGR). Still, **Figure 6B** reveals that even when using 50bp, as opposed to the 20bp usually considered, the lengths of 19% of IGRs on the minus strand and 15% of IGR on the plus strand exceeded that which is normally observed in other prokaryotes. We found that the length of the operon also had no impact on the length of the IGRs. Meaning, IGR lengths longer than 20bp were observed as frequently in operons containing two CDSs as they were in operons that were 14 to 15 CDSs in length. However, there was a preference of location for longer IGR lengths. In 44% of cases, excessive IGR lengths were between the first two CDSs of an operon. It should be noted that eight of these long IGRs were within operons containing just two CDSs. Hence, in the cases of bi-cistronic operons, we will miss entire operons if we filter and exclude CDSs based on IGR length. In fact, in many instances, these long IGRs were also between all the CDSs of longer operons. We calculated that 26% of our real operons would not be predicted if IGR length is used as a feature. This parameter was therefore excluded from this algorithm.



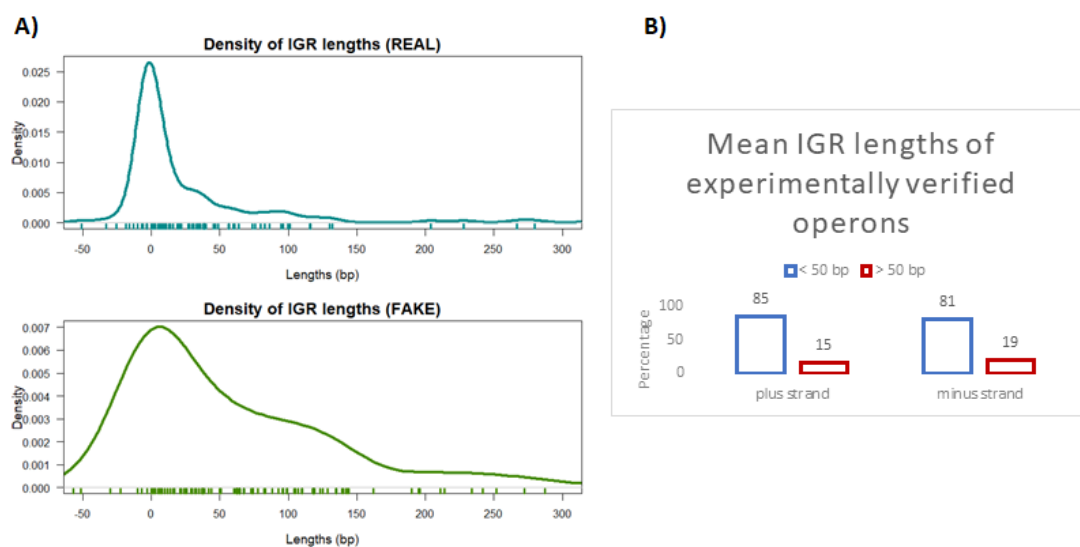


Figure 6: Comparison of intergenic distance between real and fake operons.

A) Although the IGR lengths peaked around 0 nucleotides, there were still far too many IGR lengths that were longer than what is usually observed in operons of other prokaryotes. **B)** A quarter of the IGRs on the plus strand and 19% of IGR on the minus strand exceeded 50bp. Limiting the operon predictions by IGR length would result in the exclusion of more than a quarter of verified operons (26%). This parameter was therefore excluded from the algorithm.

3.1.7 Motifs at the start or the end of an operon

Unfortunately, MEME found no consensus sequence for any of these regions. We suspect that because operons are so dynamic - i.e., their compositions change with respect to their environment - we may need to first determine which genes of an operon are always expressed – even when experimental conditions vary. That is, consensus sequences may be more easily determined for static operons. Once these static operons are known, we could repeat the search for motifs that may be responsible for delineating the boundaries of operons. We also need to be open to the possibility that these motifs may not necessarily flank operons, but that they may lie within the CDSs or even the IGRs of operons. This parameter was therefore put on hold for future consideration.

3.2 Algorithm validation

3.2.1 Multiple Linear Regression Analysis

We anticipated that the FD between adjacent CDSs would have the greatest effect on the outcome. Surprisingly, the MLR analysis showed that the greatest impact on the outcome variable was the new parameter - maximum FD between the IGR and its flanking CDSs (coefficient estimate = -0.37). Meaning that each time this parameter is **decreased** by just 1 unit, the total number of operons predicted, increases by 0.37 percent. Naturally, this number is not very high, because our verified TP list is small, so there are less operons to catch. The next most significant parameter was the minimum CDS coverage, followed by the minimum IGR cut-off. The least significant parameter was the maximum FD between adjacent CDSs. The MLR showed that all four predictor variables were highly statistically significant ($< 3.8 \times 10^{-16}$). As expected though, despite their significance, these variables accounted for 40% of the variability (adjusted $R^2 = 0.4$). We suspect that the other variable (splitting putative operons at the point where correlation breaks between distant adjacent CDSs), discussed in **Section 2.3**, may have a large impact on the outcome, but this could unfortunately not be tested. Surprisingly, several authors have argued against using R^2 as a strict predictive measure of model's performance. They argued that R^2 may be a biased, insufficient and misleading measure of predictive accuracy, and that RMSE may give a much better indication of the accuracy of a model (Alexander et al. 2015; Li 2017). We therefore measured the RMSE and the results showed that the error rate was definitely reduced in the final model (2.6), compared to the baseline model (3.4). Similarly, the MAE dropped from 2.5 to 2.1 in the final model, showing that the decision tree performs better when these four parameters were used, than if they were excluded. The pruned tree model (which combats overfitting of data) again calculated no default cutoff for a max FD between adjacent CDSs (the least significant predictor). However, it computed that certain cutoffs can be utilized to correctly predict $\geq 37/50$ EVOs for most strains.

This would be achieved if we restrict the:

- i) **5.5** < FD between CDS and flanking IGRs <=**13**,
- ii) min CDS coverage <= **7.5** and,
- iii) min IGR coverage <= **6.5**.

These values, together with our observations were considered for the default parameters of COSMO.

3.2.2 Comparison to existing algorithms

Finally, we compared the total full-length operons called by COSMO to REMap and Rockhopper. We settled on using the 10x cutoffs for REMap as per their publication, because it performed better than their algorithm's default setting of 20x. As per the REMap algorithm, this cutoff applies to both the CDS and IGR coverages.

COSMO with its four parameters as input (**Control:** min CDS = 1x; min IGR = 4x; max FD of IGR-vs-CDS = 6x; max FD adjacent CDSs = 7x. **Experimental:** min CDS = 2x; min IGR = 1x; max FD IGR-vs-CDS = 5x; max FD adjacent CDSs = 5x) was able to accurately predict more operons under both the control and experimental conditions (52% and 50%, respectively) than REMap (46% and 48%, respectively) and Rockhopper (48% in total), as shown in **Table 2**. This is significant because most existing algorithms do not generate condition-specific operons. Rockhopper for example, has only a total value, because it predicts operons based on differential expression and also does not allow user-defined cutoffs like REMap and COSMO. When the control and treated samples are submitted independently for operon predictions, Rockhopper generated identical reports.

Moreover, when the number of operons predicted under both conditions were combined, COSMO's total predicted operons (60%) also exceeded that of REMap (50%) and Rockhopper (48%). COSMO was also the frontrunner with regards to

sensitivity (88%), compared to REMap (69%) and Rockhopper (57%). This however came with the usual trade-off in precision where Rockhopper performed better (77%) than COSMO (65%) and REMap (64%). However, overall, COSMO was not only correctly identifying more operons, but it was also doing it more accurately, with F1 scores for COSMO, REMAP and Rockhopper at 75%*, 67%* and 66%*, respectively.

We are aware that the total number of operons caught may still seem low. In many other studies where operons were predicted, the performance metrics are often over 80% or 90% for sensitivity, specificity, and F1 accuracy scores (Zheng et al. 2002; Bockhorst et al. 2003; Laing et al. 2008; McClure et al. 2013). However, in these studies, operons are usually split into gene pairs. This obviously leads to a much bigger true positives (TPs) test set, allowing for more correct predictions to be made. With COSMO, if an operon consisting of 5 genes is split after CDS 2, our algorithm computes it as unpredicted, or a FN, whereas in other studies, since only one gene pair was not predicted, it will be counted as 3 TPs + 1 FN. For COSMO we chose to call the full length as the operon, since it's a true reflection of the length.

*The final F1 scores may be 1% higher than when this calculation is done because the precision and recall values were rounded off in Table 2.

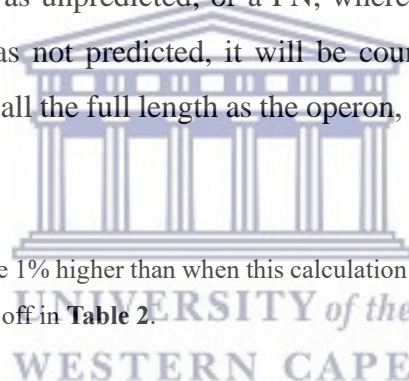


Table 2: Performance of the three operon prediction tools: COSMO, REMap and Rockhopper.

Algorithm (condition) ^a	Predicted operons Under Specific Condition ^b (%)	Total Correctly predicted Operons ^c (%)	Sensitivity (%)	PPV ^d (%)	F1 score (%)
Cosmo (ctrl)	52	60	88	65	75
Cosmo (exp)	50				
Remap (ctrl)	46	50	69	64	67
Remap (exp)	48				
Rockhopper	48	48	57	77	66

^a control (ctrl) or experimental (exp) condition

^b percentage of operons called under control (no RIF-stress) or experimental condition (RIF stress);

^c the percentage of the total number of correctly predicted operons

^d the positive predictive value.

3.2.2.1 Unique predictions

In terms of unique predictions from the list of EVOs, seven operons were predicted only by COSMO (Rv0046c-Rv0047c, Rv0096-Rv0102, Rv0287-Rv0288, Rv1964-Rv1966, Rv1966-Rv1971, Rv2743c-Rv2745c, Rv3516-Rv3517), as displayed in **Figure 7A**, and three operons were predicted by Rockhopper (Rv2877c-Rv2878c, Rv3917c-Rv3919c, Rv3921c-Rv3924c). REMap was able to predict one operon

from literature that was not predicted by COSMO or REmap (Rv3417c-Rv3423c). This is discussed in greater detail in section 3.2.2.3.

3.2.2.2 Condition-dependent mapping

Figure 7B, illustrates the ability of COSMO to distinguish between operons predicted under control conditions from those predicted when under stress (RIF treatment). Four operons were predicted solely under control conditions, while five operons were predicted as active only under RIF stress. As shown in **Figure 7C**, although REmap predicted less operons than COSMO, it also demonstrated the ability to classify operons in a condition-specific manner - a distinction Rockhopper was not able to make. The Venn diagrams were created using ‘*matplotlib-venn*’ 0.11.6 (Hunter 2007) in Jupyter Notebook 6.0.3 (Kluyver et al. 2016).



A) Venn diagram of correctly predicted operons for COSMO, REmap and Rockhopper

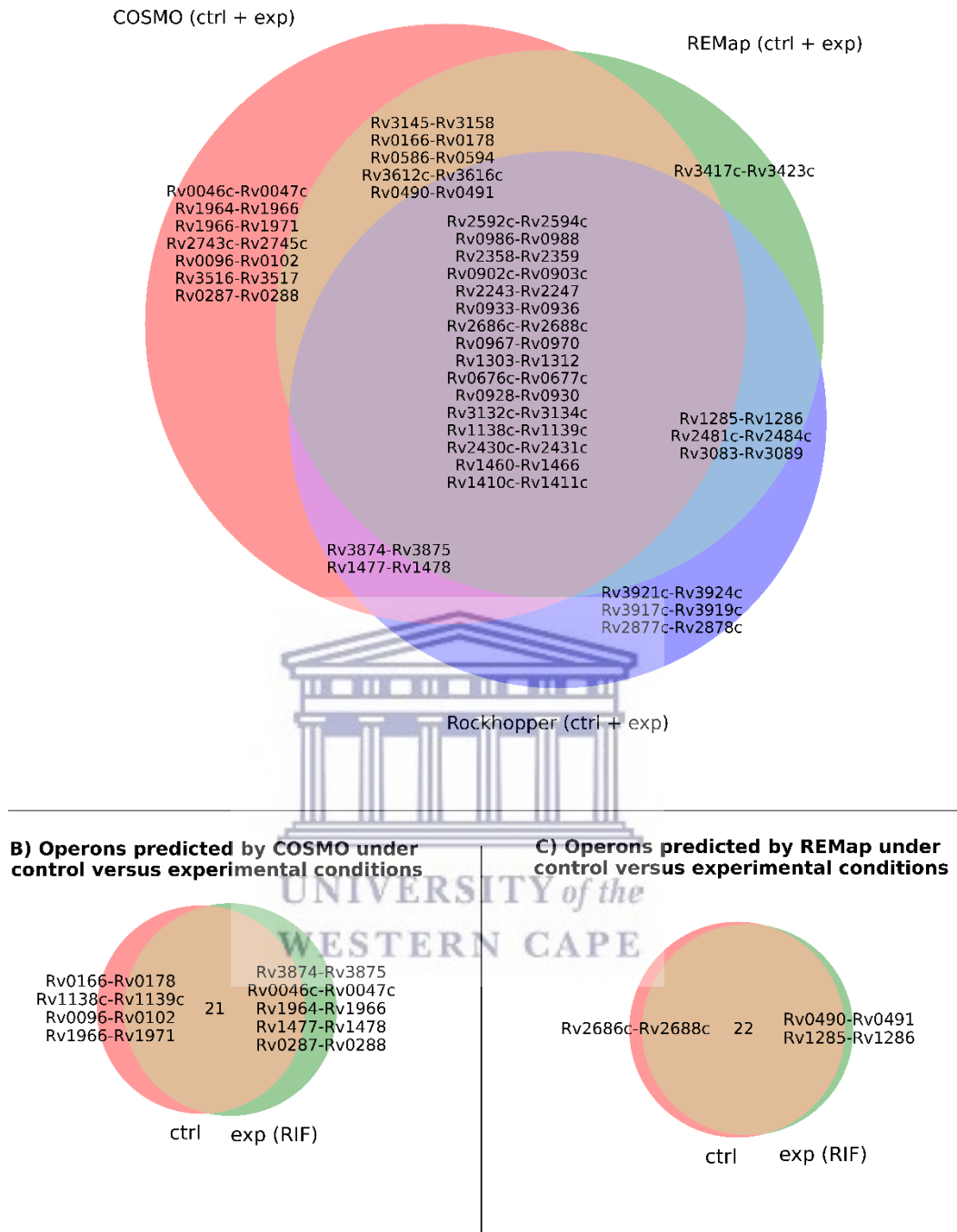


Figure 7: Comparison of operons predicted by COSMO, Rockhopper and REmap. A) The intersection of operons from COSMO, Rockhopper and REmap. Only one operon was uniquely predicted by REmap, while three operons were uniquely predicted by Rockhopper. Seven operons were predicted solely by COSMO. **B)** COSMO also predicted four operons as only expressed under control conditions, while five operons were predicted only under RIF stress. **C)** REmap was also able to distinguish two operons that were predicted under control conditions from the one specific to RIF stress. Rockhopper is not shown, because the algorithm only predicted combined differentially expressed operons.

3.2.2.3 Operons not predicted by COSMO

We investigated the operons not predicted by each algorithm, to determine whether the operons were either shorter than those reported in literature (FNs), or longer (FPs). **Table 3** shows the 50 EVOs and their prediction calls across the three algorithms. Only 8% (n = 4) of the operons incorrectly predicted by COSMO were shorter than those in literature - or could be considered as FNs; in contrast to REMap and Rockhopper for which 22% (n = 11) and 36% (n = 18) of operons were FNs, respectively. Most of the operons not found by COSMO were predicted to be slightly longer than those in the literature and none of the operons were completely unexpressed (zero genes expressed). On the contrary, REMap and Rockhopper called all the genes of some operons as unexpressed (n = 5 and n = 6 respectively). Although DOOR 2.0 no longer predicts operons, we compared our results to a list of operons already predicted by DOOR 2.0 for *Mtb* (not shown). COSMO also outperformed DOOR, by predicting nearly twice as many operons as DOOR, which correctly predicted only 16 of the total operons (32%).

One especially interesting feature of COSMO is that it is able to predict operons with CDSs that are expressed at very low levels. As briefly discussed in section 3.2, the expression levels of 5x and below, for both the CDS and IGR, resulted in the highest number of total predicted operons for most isolates. This is because COSMO is able to bypass low expression levels, while rather taking advantage of a maximum FDs between CDSs and between an IGR and its flanking CDSs. This is also one of the reasons why REMap and Rockhopper were not able to predict some operons from literature. One example of this is the operon Rv3516-Rv3517, which was predicted by COSMO, but not by REMap or Rockhopper, since its expression levels were very low. Therefore, if the expression of an operon is deemed detrimental by *Mtb* for its virulence or survival or it's biologically redundant, and it deliberately downregulates the expression of this operon, COSMO would still be able to predict those operons and record the downregulated expression.

Table 3: EVOs predicted by three algorithms.

Operon	COSMO	REMap	Rockhopper
Rv0046c- Rv0047c	√	<i>Rv0043c-Rv0048c</i>	NOT EXPRESSED
Rv0096-Rv0102	√	Rv0099-Rv0101, Rv0102	Rv0096-Rv0101
Rv0166-Rv0178	√	√	Rv0167-Rv0178
Rv0287-Rv0288	√	<i>Rv0280 – Rv0291</i>	<i>Rv0287-Rv0289</i>
Rv0490-Rv0491	√	√	NOT EXPRESSED
Rv0586-Rv0594	√	√	Rv0586-Rv0589, Rv0591-Rv0594
Rv0676c- Rv0677c	√	√	√
Rv0735-Rv0736	<i>Rv0735 - Rv0737</i>	Rv0732-Rv0735	<i>Rv0732-Rv0736</i>
Rv0902c- Rv0903c	√	√	√
Rv0928-Rv0930	√	√	√
Rv0933-Rv0936	√	√	√
Rv0967-Rv0970	√	√	√
Rv0986-Rv0988	√	√	√
Rv1138c- Rv1139c	√	√	√
Rv1161-Rv1164	<i>Rv1161 Rv1166</i>	<i>Rv1161 – Rv1165</i>	<i>Rv1161-Rv1166</i>
Rv1285-Rv1286	<i>Rv1284 – Rv1289</i>	√	√
Rv1303-Rv1312	√	√	√
Rv1334-Rv1336	<i>Rv1331 – Rv1341</i>	<i>Rv1331 – Rv1341</i>	NOT EXPRESSED
Rv1410c- Rv1411c	√	√	√
Rv1460-Rv1466	√	√	√
Rv1477-Rv1478	√	<i>Rv1476 – Rv1481</i>	√

Table 3: EVOs predicted by three algorithms. (continued).

Rv1483-Rv1484	Rv1483 – Rv1485	Rv1483 – Rv1485	Rv1483-Rv1485
Rv1660-Rv1661	<i>Rv1659 – Rv1665</i>	<i>Rv1659 – Rv1661</i>	Rv1661-Rv1664
Rv1806-Rv1809	Rv1806 - Rv1807, Rv1808 - Rv1809	Rv1807, Rv1809-Rv1811	NOT EXPRESSED
Rv1826-Rv1827	<i>Rv1825 – Rv1829</i>	<i>Rv1821 – Rv1832</i>	Rv1822-Rv1826, Rv1827-Rv1828
Rv1908c- Rv1909c	<i>Rv1907c – Rv1909c</i>	<i>Rv1907c – Rv1909c</i>	<i>Rv1907c-Rv1909c</i>
Rv1964-Rv1966	√	NOT EXPRESSED	<i>Rv1964-Rv1975</i>
Rv1966-Rv1971	√	NOT EXPRESSED	*
Rv2243-Rv2247	√	√	√
Rv2358-Rv2359	√	√	√
Rv2430c- Rv2431c	√	√	√
Rv2481c- Rv2484c	<i>Rv2481c - Rv2485c</i>	√	√
Rv2592c- Rv2594c	√	√	√
Rv2686c- Rv2688c	√	√	√
Rv2743c- Rv2745c	√	<i>Rv2742c - Rv2745c</i>	Rv2742c-Rv2744c
Rv2871-Rv2875	Rv2871 – Rv2874, Rv2875 – Rv2876	Rv2871 – Rv2872, Rv2873 – Rv2876	Rv2871-Rv2872, Rv2875-Rv2876
Rv2877c- Rv2878c	<i>Rv2877c - Rv2883c</i>	<i>Rv2877c - Rv2883c</i>	√
Rv2931-Rv2938	<i>Rv2930 - Rv2939</i>	<i>Rv2928 - Rv2939</i>	<i>Rv2930-Rv2938</i>

Table 3: EVOs predicted by three algorithms. (continued).

Rv2958c- Rv2959c	<i>Rv2958c - Rv2960c</i>	<i>Rv2958c - Rv2960c</i>	NOT EXPRESSED
Rv3083-Rv3089	Rv3083 - Rv3085, Rv3086 - Rv3089	√	√
Rv3132c- Rv3134c	√	√	√
Rv3145-Rv3158	√	√	Rv3145-Rv3151, Rv3152-Rv3158
Rv3417c- Rv3423c	Rv3417c- Rv3418c, Rv3419c- Rv3423c	√	Rv3417c-Rv3418c, Rv3419c-Rv3423c
Rv3493c- Rv3501c	<i>Rv3492c - Rv3503c</i>	<i>Rv3492c - Rv3503c</i>	Rv3492c-Rv3501c
Rv3516-Rv3517	√	Rv3516	NOT EXPRESSED
Rv3612c- Rv3616c	√	√	Rv3612c-Rv3614c
Rv3793-Rv3795	<i>Rv3788 - Rv3798</i>	Rv3792-Rv3793, Rv3794- Rv3796	<i>Rv3789-Rv3796</i>
Rv3874-Rv3875	√	NOT EXPRESSED	√
Rv3917c- Rv3919c	<i>Rv3916c - Rv3924c</i>	NOT EXPRESSED	√
Rv3921c- Rv3924c	*	NOT EXPRESSED	√

A tick mark (√) shows that the algorithm found an exact matching operon to that in the EVO list. Operon names written in **bold** indicate that the algorithm predicted the operon to be shorter than the EVO, while *italic font* means it was longer than the EVO. That is, the operon was extended either upstream, or downstream or in both directions. NOT EXPRESSED indicates that none of the genes of the operon were expressed by that algorithm. An asterisk represents an operon that was predicted to be overlapping with the previous operon.

3.2.2.4 Composition of predicted operons

Lastly, COSMO's potential to predict novel operons was also analyzed. Under control- and RIF stress conditions, the number of putative operons and the number of CDSs forming part of putative operons were on average not notably different across samples. **Table 4** shows the average results for the 12 samples. Under control conditions, COSMO predicted approximately 71% of the 4109 protein coding genes in *Mtb* to be constituents of operons (n = 952 operons), compared to 72% under RIF stress. Fewer genes were expressed as single genes under RIF (n = 1035), compared to control conditions (n = 1045), but there were also less unexpressed single genes under RIF-stress (n = 119) versus under control conditions (n= 156). This was because most of the genes that were expressed under RIF, were expressed in operons (n = 2960 genes within 965 operons). Interestingly, longer operons were more common under control conditions and the largest operon was predicted under control conditions, consisting of 15 CDSs. The results for all the operons can be found on our GitHub page at:

https://github.com/SANBI-SA/COSMO/blob/master/Supplementary_data/operons_vs_genes_vs_unexpressed_per_isolate.xlsx.

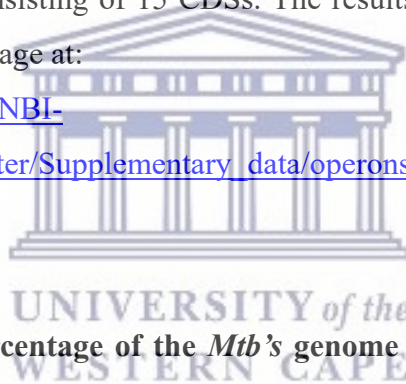


Table 4: Average percentage of the *Mtb*'s genome predicted as operons and single genes across the 12 samples.

		Control		RIF treatment	
		Number	%	Number	%
Operons	Count	952		965	
	Genes	2908	71	2960	72
Single genes	Expressed	1045	25	1030	25
	Unexpressed	156	4	119	3
Total genes		4109		4109	

4 Discussion

COSMO takes in RNA sequence reads (.bam files) together with a GTF file and identifies the operons active under varied experimental conditions. For this study, we used RIF-treatment as our experimental condition and no RIF treatment as the control. The user can provide four separate parameters as cutoffs. Although other RNA-seq based algorithms also allow users to define an expression level cut-off, to the best of our knowledge this is the first algorithm predictor that allows the user to provide a separate input for IGR coverage and CDS coverage. It is also the first operon predictor where users can define the maximum FD between adjacent CDSs as well as the maximum FD between an IGR and its flanking CDSs. Additionally, COSMO not only considers correlation of expression between directly adjacent genes, but between all CDSs within a predicted operon.

Allowing the user to provide the FDs is especially valuable. These parameters can take pre-eminence over just considering the min CDS or IGR expression levels as predictive indicators. This tolerates expression levels of CDSs and IGRs that are very low, which may be crucial for identifying operons that may have been downregulated by the bacteria in response to stress.

Our results therefore supported the findings of REMap, that a maximum IGR length may not be a useful feature for operon prediction in *Mtb* (Pelly et al. 2016) – despite being the most reliable indicator of an operon in many operon algorithms. In fact, approximately a quarter of our 50 EVOs would not have been predicted if we excluded operons based on short IGR distance. Similarly, our MLRA showed that even coverage between two adjacent CDSs, which is universal in other algorithms, and is based partly on the definition of an operon, was the least significant parameter. It was outweighed by three of our parameters – the greatest of which was the new parameter - maximum FD between IGRs and their adjacent CDSs (Dam et al. 2007; Taboada et al. 2010, 2018; Tjaden 2019; Krishnakumar and Ruffing 2022).

COSMO can distinguish between operons predicted under control conditions and operons predicted under experimental stress conditions. Being able to match which operons are expressed under each condition is valuable, because we may observe

which pool of genes are specifically expressed under one condition, while being deemed either detrimental or at the very least, redundant under another condition. This capability may eventually allow us to better understand how *Mtb* is able to circumvent and thrive under stress conditions by tailoring a condition-specific pool of genes within the boundaries of an operon, as well as which operons are never altered, but may be considered as “housekeeping operons”.

COSMO was able to predict all experimentally validated operons from literature, with 60% of the operons being exact matches to those obtained from literature, while Rockhopper predicted 48% and REMap predicted 50%. COSMO also obtained a greater accuracy, with an F1 score (75%), compared to REMap (67%) and Rockhopper (66%). With regards to those that were not exact matches; some operons were slightly shorter than in literature. However, most operons that were not identical to those from literature, were predicted to be slightly longer. On the contrary, REMap and Rockhopper not only predicted more shorter operons (FN), but they also predicted entire operons as unexpressed.

Although matching only 60% of operons from literature correctly may seem rather low, it would be more alarming if we predicted everything or if we predicted close to 100% of operons. This would contradict our understanding that different operons are expressed under different environmental conditions. The lower value compared to other operon prediction studies, was also as a result of predicting full-length operons. This is a more accurate representation of what is happening in the context of operons, than in other studies where gene pairs are predicted (Overbeek et al. 1999; Ermolaeva et al. 2001; Zheng et al. 2002; Bockhorst et al. 2003; Laing et al. 2008; McClure et al. 2013; Tjaden 2020b). For this study, we have also only tested COSMO on isolates under control versus RIF stress conditions. This was because the main objective was just to evaluate COSMO against existing operon-predicting algorithms.

COSMO was in agreement with the results that REMap and Rockhopper previously showed, in that at any given point in time, the larger proportion of *Mtb* genes/CDSs, are not operating independently, but they are instead predicted to be constituents of operons. COSMO predicted that $\geq 71\%$ of *Mtb*'s protein coding genes may constitute operons – whether it was under control or RIF-stress conditions. REMap

reported this number to be just under 60% for *Mtb* (Pelly et al. 2016), while Rockhopper predicted this as a range between 38% for *Caulobacter vibrioides* to 80% for *Vibrio cholerae* (Tjaden 2019).

This may indicate that *Mtb* and other prokaryotes have a heavy reliance on forming operons as a means of regulating its genome. We also predicted that under RIF stress, operons seemed to form more frequently and shorter operons were more common under RIF stress. This may suggest that under RIF-stress, *Mtb* activates operons on an *ad hoc* basis, to swiftly and efficiently handle the adversity it faces at that specific time. The significance of this, is that *Mtb*, the species responsible for causing Tuberculosis, is known for its impressive ability to evade and survive within their hosts (Namouchi et al. 2016). This pathogen, which is responsible for more deaths than any other infectious agent, worldwide (World Health Organization 2019), has co-evolved with its hosts over several millennia and has continuously outsmarted the myriad of drugs that were carefully designed to disrupt its virulence at a gene or SNP level (Hoagland et al. 2016; Coll et al. 2018). If *Mtb* favours operons under stress conditions, then it may make more sense to study its evasive tactics in the context of operons, rather than by looking at mutations or differential expression of individual genes – which is how it was traditionally done. However, further analyses would have to be carried out to determine whether creating shorter operons allows it to have a tighter control over gene regulation or whether it has an alternative purpose. One of our current analyses involves taking a deeper look at the functions of the genes in the operons that are changing and differentially expressed under each experimental condition.

5 Limitations and Future work

One of the limitations we had in this study was the very small list of validated operons. Once this list becomes more populated, we may be able to evaluate the algorithm's accuracy using more traditional methods such as sensitivity and specificity ROC curves. We are hoping that once this algorithm is tested across a variety of lineages and experimental conditions, we may be able to detect the static CDSs of operons. This may aid us in further optimizing the algorithm if static operons can lead us to one or several consensus motifs that can be used as a feature or parameter in the algorithm design. Lastly, an experimental validation will have to ensue on carefully selected candidate operons predicted by COSMO to further gauge its performance.

This analysis will also be extended to other *Mtb* families and to *Mtb* genomes exposed to different environmental conditions. This should generate a higher number of matches to operons published in literature, since the current experimentally validated operon list we used, consisted of operons discovered from a variety of different *Mtb* lineages and from a variety of experimental conditions. However, COSMO has already demonstrated an improved capacity to identify existing operons when compared to REMap and Rockhopper. Additionally, because it does not rely on inherent *Mtb*-specific traits for operon prediction, it could also be utilized for operon predictions in other microorganisms.

6 Data Availability Statement

The datasets and scripts generated/analyzed for this study can be found in the links below.

COSMO algorithm: <https://github.com/SANBI-SA/COSMO>

GTF and other coordinate files: https://github.com/SANBI-SA/COSMO/tree/master/Algorithm_parameter_testing/GTF_%26_other_coordinate_files

Some wiggle files: https://github.com/SANBI-SA/COSMO/tree/master/Algorithm_parameter_testing/Wiggle_files

Calculating total correctly predicted operons: https://github.com/SANBI-SA/COSMO/blob/master/Algorithm_parameter_testing/Python_scripts/calculate_total_correct_operons.py

Coverages - Genes and IGRs: https://github.com/SANBI-SA/COSMO/blob/master/Algorithm_parameter_testing/Python_scripts/average_operon_genes_and_IGRs.py

Coverages - UTRs: https://github.com/SANBI-SA/COSMO/blob/master/Algorithm_parameter_testing/Python_scripts/average_coverage_UTRs_operons.py

Calculating TP, TN, FPs: https://github.com/SANBI-SA/COSMO/tree/master/Algorithm_parameter_testing/python_scripts_for_prediction_calls

MLR and decision tree script: https://github.com/SANBI-SA/COSMO/tree/master/Algorithm_parameter_testing/R_script

CHAPTER 4

GENOTYPE- AND CONDITION-SPECIFIC OPERON PREDICTION FOR *MYCOBACTERIUM TUBERCULOSIS* UNDER RIFAMPICIN STRESS

Abstract

Background: Bacteria often form operons in response to unfavourable environmental conditions. Operons are sets of adjacent genes which are co-expressed as a single polycistronic mRNA. They are not just dynamic in terms of their length, but several studies have shown evidence of up- or downregulation of entire operons, when exposed to stresses. We previously showed that our algorithm, COSMO, outperformed the best operon predictors and was also able to better distinguish between operons predicted under control conditions versus RIF stress. In this study, we aimed to see if operons were differentially expressed and if their lengths were altered under RIF stress, and whether these modifications occurred in a genotype-specific or strain-specific manner. We also aimed to understand what the biological implications of operon modifications could be.

Methods: Using COSMO, we predicted operons for 64 *Mtb* samples from RIF-resistant lineage 2 and lineage 4 strains, as well as for drug sensitive wild type strains. Predicted operons were evaluated against a set of 50 experimentally verified operons (EVOs). Operon expression changes as well as changes in operon lengths were predicted under RIF-stress conditions, for each genotype. Lastly, we predicted operons under hypoxia stress using publicly available RNA seq datasets of nine strains.

Results: We predicted 70% of the full-length EVOs across the genotypes and our sensitivity, precision and F1 accuracy scores showed significant improvements. A total of 32% of operons maintained the same length as the EVOs, even when exposed to RIF stress. These operons may be under selection pressure and could possibly serve as housekeeping operons, which would be interesting targets for anti-tubercular drugs. Only one operon, Rv0676c-Rv0677c, also known as the MmpS5-MmpL5 efflux system, was significantly downregulated in the *rpoB* mutant

genotype ($\log_2 \text{FC} \geq 0.58$; $p < 0.05$). *Mtb* seemed to prefer to alter the length of an operon under RIF stress, over up- or downregulating the expression levels of an entire operon. Still, even for the operons which lengths were modified, most strains (80%, $p = 1.4 \times 10^{-9}$) regulated their operon lengths in genotype-specific manner, rather than for each strain to individually modify its operon length in response to RIF-stress. Proteins involved in lipid metabolism were the most frequent targets where operons were split to produce shorter operons. Regulatory proteins were favoured for creating operons that were longer than the EVOs, and proteins involved in ATP-related processes were under the most intense positive selection pressure to constitute static TP operons. Finally, an additional operon, which was not previously predicted by COSMO under RIF-stress, was predicted for strains grown under hypoxia. This operon was confirmed to participate in the hypoxia pathway.

Conclusion: In this analysis COSMO was able to i) correctly identify more EVOs using different genotypes, ii) demonstrate that *Mtb* operons generally resisted being reorganized, and resisted being up- or downregulated under RIF stress with respect to their genotype, iii) distinguish between operons predicted under RIF stress from those predicted under hypoxia stress and v) show that the operon predictions may help us to assign meaningful biological inferences relating to the pathogen's adaptation to stress.



UNIVERSITY of the
WESTERN CAPE

1 Introduction

The *Mycobacterium tuberculosis* complex (MTBC) consists of seven main human-adapted lineages, which are all obligate pathogens. There are approximately 1200 single nucleotide polymorphisms (SNPs) which separate these MTBC strains into their distinct lineages. Although the number of classifying SNPs may seem few, they are still enough to produce a notable phenotypic difference in the way each *Mtb* lineage metabolizes lipids, in their degree of virulence, resistance and immunogenicity (Coscolla and Gagneux 2014).

Lineages 5 and 6 (L5 and L6) belong to the *Mycobacterium africanum* species. The other five lineages belong to the respiratory system pathogen, *Mycobacterium tuberculosis* (*Mtb*). *Mtb* is the causal agent behind tuberculosis. Before the SARS CoV-2 pandemic, it was the most infectious and the most fatal global disease. Its ancient lineages, L1 (Indo-Oceanic) and L7 (Ethiopia), are geographically confined strains. In contrast, the more recently evolved lineages consist of L2 to L4, also known as the Beijing, East-African Indian and Euro-American lineages, respectively. These newer lineages have evolved with virulent traits which favour transmissibility and are therefore more prevalent, as evident by their global presence (Comas et al. 2013, 2015; Nebenzahl-Guimaraes et al. 2016; Orgeur and Brosch 2018). However, despite having evolved more recently, L2-L4 already demonstrate significant differences between them. An *in vitro* study by Ford et al. (2013) showed that due to a higher mutation rate, the L2 lineage acquired drug-resistance more rapidly than the L4 lineage.

A 2006 study by Gagneux *et al.*, showed that *Mtb* strains from different lineages demonstrated distinct levels of resistance and fitness costs when treated with the same dose of rifampicin (RIF), depending on the location of the mutation. In general, mutations in single drug resistance (DR) genes have been extensively studied to observe how they fluctuate across lineages.

For example, Ford et al., (2013) showed that L2 strains have mutations in the *katG* and *inhA* genes, which confer a resistance to isoniazid (INH). Likewise, several

studies also confirmed the link between differential expression of single DR genes and *Mtb* virulence (Manganelli et al. 2004b; Lam et al. 2008; Garima et al. 2015).

However, many DR strains harbour no mutations in any of these DR genes (Chan et al. 2007; Al-Saeedi and Al-Hajoj 2017). Moreover, research has shown that drug resistance in some strains were only conferred when several DR genes were mutated. More importantly, some DR phenotypes were only observed when **all the** genes which make up an **operon** were mutated, or if the expression of the entire operon was **up- or downregulated** (Banerjee et al. 1994; Bretl et al. 2012; Hunt et al. 2012).

An operon is a set of genes which are transcribed as a single poly messenger RNA (Price et al., 2006). They are mostly prevalent in bacteria and archaea, although some operons have been identified in eukaryotes, such as nematodes and the fruit fly (Spieth et al. 1993; Brogna and Ashburner 1997). Operons often form in response to changing environmental conditions, to aid microbes to rapidly and efficiently respond to environmental stresses (Zaidi and Zhang, 2017). This could explain why the majority of *Mtb*'s genes are expected to not function independently. Previous studies predicted that ~60% of the *Mtb* genome may operate within these higher module operon structures (Pelly et al., 2016).

However, despite the prevalence of operons and their association with drug resistance, to the best of our knowledge, no other studies have reported on how operons adapt across lineages/genotypes in control versus rifampicin (RIF) stress conditions.

We used our operon prediction tool, COSMO (<https://github.com/SANBI-SA/COSMO>) to predict how operons are reorganized across genotypes. COSMO was previously compared to three of the best performing operon predictors. It was shown to outperform all competitors in the number of correctly predicted full-length operons, and in the accuracy of these predictions.

We therefore used COSMO to generate an overview of operons which changed (dynamic operons) or maintained their lengths (static operons) under RIF stress, using strains from two lineages (L2 and L4). We also used *limma voom* to observe whether the entire operon's expression levels were altered in response to RIF stress with respect to each genotype. Lastly, we observed if a different set of operons could be predicted under hypoxia stress.

2 Methods

We previously predicted operons with COSMO for Lineage 2 Beijing strains of both low and high MIC_{RIF} statuses. In this study, we increased the sample size by adding additional strains from lineage 2 and 4 and predicted operons per genotype. A genotype refers to the *Mtb* family with its MIC status. That is, a genotype could be Family X low MIC or wild type (WT) strains or *rpoB* mutants, etc. All strains were RIF resistant, except for the WT strains, which were RIF sensitive.

By using these additional strains, we aimed to observe if:

- i) COSMO predicted more EVOs with a more **varied** collection of **strains**.
- ii) any operons remained the **same length** (static) across the **genotypes**. If yes, were they static true positives (TPs), static false positives (FPs) or static false negatives (FNs).
- iii) any operons **changed** in **length** (dynamic) under RIF stress and if so, whether they changed in a genotype-specific or more strain-specific (individual) way.
- iv) any operons were up- or downregulated under RIF stress.
- v) a functional annotation (FA) of these predictions could help us infer some of the possible biological reasons for the changes in the operons, and
- vi) unique operons can be predicted under **hypoxia stress**.

a) Sample Collection

The samples were obtained from three *Mtb* patients and RNA-sequenced (ethics number N10/04/126). The lineages and drug resistance profiles were confirmed using the *TBProfiler* tool (Phelan et al. 2019) in Galaxy (<https://galaxy.sanbi.ac.za/>) (Jalili et al. 2020). Except for the WT genotype, all other strains were RIF resistant. As shown in **Figure 8**, the 64 samples were from lineage 2 and lineage 4. L2 consisted of three wildtype (WT) biological replicates (BR) which were grown **only** under control conditions (no RIF treatment). It also consisted of three BRs of **high** MIC of 150 ug/ml, from the Beijing genotype, grown under control conditions and three BR grown under experimental conditions (RIF treatment). Similarly, three **low** MIC (40 ug/ml) BRs from the Beijing genotype were grown under control conditions and three under RIF stress. Under L4 we grew three BRs which were *rpoB* mutants under control conditions and three BRs under RIF-stress. Furthermore, L4 also consisted of three **high** MIC (150 ug/ml) BRs and three **low** MIC (40 ug/ml) BRs belonging to the Family X genotype, grown under control conditions and under RIF stress. Each BR of the Family X samples contained at least three technical replicates (TR). Cultures were grown in 7H9 media until mid-log phase and exposed to RIF for 24 hours. Both the high- and low MIC strains were exposed to a quarter MIC of RIF.

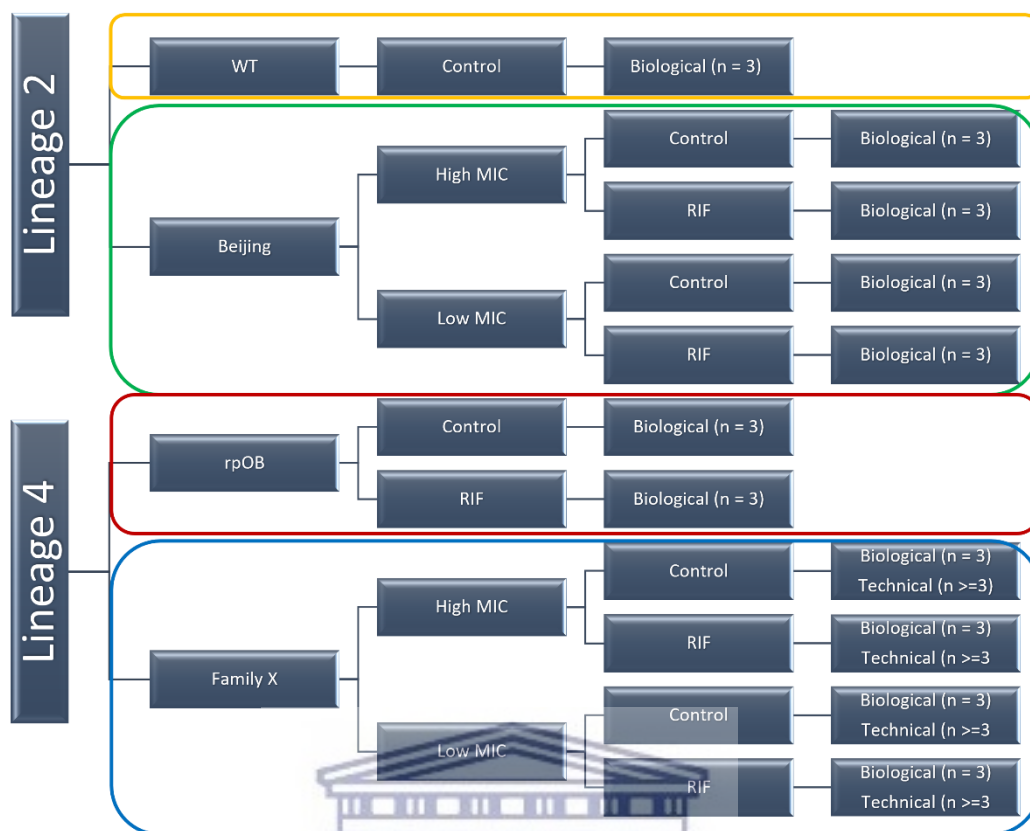


Figure 8: Outline of the study design for lineage 2 and 4 samples. A combined 64 samples were used in this study which spanned Lineage 2 and Lineage 4. At least three biological replicates were used for each of the four families, namely: WT, *rpoB* mutants, Family X and Beijing. Each Family X BR also had at least three TRs. A genotype was considered to be the family together with its MIC status. Therefore, while the WT strains was one genotype and the *rpoB* mutants were another, the Beijing and Family X families consisted of 2 genotypes each. Thus, there is a total of 6 genotypes in the figure above.

b) RNA extraction and sequencing

RNA extraction was carried out using the FastRNA Pro Blue kit (MP Biomedicals, Germany) and residual DNA was treated with DNase (Promega, WI, USA). Ribosomal depletion was performed with the bacterial option as probes for hybridisation of rRNA (TruSeq Total RNA, USA). Primer design and RNA-seq were carried out at the Agricultural Research Council (ARC) sequencing facility in Pretoria, South Africa, using the TruSeq DNA and RNA CD Indexes (I7 and I5 adapters) and the Illumina HiSeq 2500. The strand-specific protocol was confirmed as the fr-firststrand library type, using the *RSeQC v2.6.4* ‘Infer Experiment’ tool (Wang et al. 2012a).

All fastq files have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE203032.

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE203032>

c) Trimming and alignment

A FastQC (Andrews 2022) check in Galaxy (<https://galaxy.sanbi.ac.za/>) showed that the data was of high quality (mean PHRED > 30), but the reads were nonetheless trimmed and adapter sequences were removed using Trimmomatic V.0.38 with a PHRED = 20 with a sliding window of 4 (Bolger et al. 2014). Reads were aligned to H37Rv (NC_00096.3) using BWA-MEM 0.17.1 (Li and Durbin 2009). The quality of the bam files were checked using Samtools 1.9 (Li et al. 2009), to make sure ~90% of reads were paired, ~90% of reads were aligned to the reference genome, and that the average size of the reads were ~200bp, according to the RNA-seq analysis best practices (Conesa et al. 2016).

2.1 Predicting operons

For each strain, the RNA-seq bam files were submitted to COSMO, using its default settings. The predicted operons were compared to a list of 50 EVOs and grouped into the prediction calls: TPs, FPs and FNs, using a custom python script, where:

- i TP operons were the same length as the EVO
- ii FP operons were longer
- iii FN operons were shorter.

We could not classify true negatives (TNs) in the absence of a set of genes confirmed as never forming part of an operon.

As shown in **Table 5**, for a biological replicate (BR) to be assigned a call (TP, FP or FN), at least 2/3 of its technical replicates had to have the same call. Then for the genotype to be assigned the call, all three biological replicates (BR) had to reach a consensus call. If these criteria were not met, then the prediction was considered a heterogeneous call.

Table 5: Three scenarios demonstrating the rules for assignment base on replicates.

	BR_1 ^a			BR_2 ^b			BR_3 ^c			Call
Scenario 1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Scenario 2	✓	X	✓	✓	✓	✓	✓	✓	X	✓
Scenario 3	✓	✓	✓	✓	✓	✓	✓	X	X	X

^abiological replicate 1

^bbiological replicate 2

^cbiological replicate 3

Each column contains at least three technical replicates per biological replicate.

2.1.1 Total number of operons predicted

For the first part of our analysis, we tested if we were able to predict more EVOs when we use strains belonging to two different lineages, and with different MIC statuses and different drug resistance profiles (drug sensitive versus RIF resistant). We previously showed that with a small subset of nine Beijing samples, we were able to achieve a sensitivity, precision and F1 score of: 75%, 65% and 75%, respectively. Here, we repeated the evaluation to see if there were any improvements with this larger, more varied sample size. The performance metrics: precision/positive predictive value (PPV), recall/sensitivity, and F1 score, were calculated as:

$$Precision/PPV (\%) = \left(\frac{TP}{TP + FP} \right) * 100$$

$$Recall/Sensitivity (\%) = \left(\frac{TP}{TP + FN} \right) * 100$$

$$F1 \text{ score} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

2.1.1.1 Sample variance

We also calculated the variance across biological samples. For the BRs, we asked how many BRs shared the same prediction. A three meant that all samples had a consensus prediction, a two meant that two samples shared a prediction and a one meant that none of the samples agreed on the prediction.

The TRs were a bit more complicated, since some had four or five replicates. Scores were between one and five. In the case where there was only three TRs, the scoring system was the same as for the BRs. In the case of four replicates, a score of three and four still represented the number of samples which had the same prediction. However, to receive a two, two samples must have had the same prediction, while the other two also had to have the same prediction. In the case where two were the same, and one prediction fell into one category and the other in another, the sample received a score of one.

We used the coefficient of variation (CV), which measures the ratio/percentage of standard deviation to the mean. This gives an indication of how reliable/variable our predictions were across the samples. A value above 100%, indicates that the variability of the predictions was greater than the mean, and the predictions were therefore volatile. A value below 100% means that there is less variability and that the results were more reproducible (Shechtman 2013; Khaw et al. 2019).

The coefficient of variance was calculated as:

$$CV (\%) = \frac{\sigma}{\text{mean}} \times 100$$

The calculations were carried out using Microsoft Excel's V2301 built-in functions.

2.1.2 Overall static operons

Although COSMO is a prediction tool, and therefore, predicted operons are not final proof of operon organizations, COSMO is based on RNA expression data. It therefore still allowed us to track changes in gene expression and compare fold changes between adjacent genes and IGRs. We could therefore also make certain inferences from the way prediction calls fluctuated with respect to genotype for the distinct conditions under which strains were grown. In this part of the analysis, we aimed to observe whether certain operons received the same prediction call (**static**) across genotypes, and experimental conditions. We were firstly interested in **static TP** operons. That is, operons which were predicted to be the exact lengths as the EVOs and of which the TP prediction never changed under RIF stress. We reasoned that if these **TP** operons are **never** reorganized – even under RIF-stress, then they may be good test cases for **finding motifs**, which could potentially be used to improve COSMO's prediction rate. More importantly, static TP operons may indicate that these operons are under selection pressure, lending evidence to the hypothesis that just as there are established housekeeping genes, there may also be housekeeping operons (Naville and Gautheret 2009). Static TP operons may also be useful in guiding us to the most impactful promoters, hub proteins or biochemical pathways, which would in turn allow for more potent drug targets.

For the second part of the analysis, we scrutinized operon predictions which were static FP or FN across the genotypes. We reasoned that static FP predictions (operons which were ALWAYS longer than the EVOs) may represent important, undiscovered genes/biological processes (BPs) required under RIF stress. Similarly, static FN predictions (consistently shorter operons) could mean that the genes which were not predicted to be part of the operon may either be detrimental, impose a too large of a fitness cost, or that they may be redundant.

2.1.3 Dynamic operons

After identifying possible static operons, we then analyzed the operons which changed in length (dynamic) under RIF-stress across the genotypes. To examine these operon reorganizations, we split the analysis in two.

- i) The aim of the first part of the analysis was just to ascertain if there was any operon length modification under RIF stress. We argued that if a genotype predicted an operon to be one call, for example, **TP** under control conditions, then whether the operons were a consensus **FN** under RIF stress, or a heterogeneous call consisting of **TP/FP/FN** calls, then technically in both cases the operon lengths changed under RIF stress.

- ii) For the second part of the analysis, we made a distinction between the two outcomes. Here we reported whether changes in operon lengths under RIF stress were genotype-specific (consensus calls among strains) or strain-specific (heterogeneous calls) under RIF stress. Here we were not just interested in whether we were observing operon changes under RIF stress, but we wanted to observe whether strains responded similarly to the changing environment, as if they received a genotype-specific signal on how to modify the operon length, or whether each strain behaved autonomously to circumvent the RIF stress.

The paired t-test was used to evaluate statistical significance.

2.2 Differential expression of operons

There are two ways in which an operon can be modified under RIF stress; either by changing the length of the operon or by altering the expression levels of the entire operon. Naturally, if the expression levels of individual genes of an operon are altered independently, it may result in a lack of co-expression of the adjacent genes. This would automatically cause a change in the operon length. For example, the

operon may split at the gene where the respective co-expression breaks the correlated gene expression.

The second way in which an operon may be modified, is when the expression levels of all the genes of an operon are altered, but with respect to each other (co-expression/coregulation). In this case the **length** of the operon will **remain the same**, but the overall expression will be upregulated or downregulated, resulting in a differentially expressed operon (DEO). So, in this part of the analysis we examined whether the average operon expression levels (coverages) were up- or downregulated for each of the genotypes under RIF-stress (DEO).

This was done using *limma voom* 3.48.0 (Smyth 2005; Law et al. 2014; Liu et al. 2015) in Galaxy (<https://galaxy.sanbi.ac.za/>). We created a GTF file which contained the start and end coordinates of the entire operon, as opposed to the usual gene coordinates captured in a GTF file. The DEO analysis therefore **included** the intergenic regions between operon genes. We used the “with sample quality weights” parameter to deal with outliers by downweighing them (Ritchie et al. 2006). The Trimmed mean M-values normalizations (TMM) was applied, where a weighted trimmed mean of the log expression ratios is used to scale the counts for the samples (Robinson and Oshlack 2010). Finally, we filtered the result using the t-tests relative to a threshold (TREAT) method, which is a robust method that combines the FDR and log₂FC values to analyze differentially expressed data in both a statistically and a biologically significant manner (McCarthy and Smyth 2009). We used the parameters $\log_2(1.1) = 0.13$, and the adjusted p-value or false discovery rate (FDR) of <0.05 , based on the Benjamini-Hochberg method (Benjamini and Hochberg 1995). This initial low log₂FC was suggested for the TREAT method, because TREAT is designed to constantly adjust the FDR as it ranks the genes/operons, which could cause most of the operons to be excluded if the initial FC value was any higher (false negatives).

FC cut-off values are usually arbitrary and are chosen by the researcher. For example, Miryala et al., (2019) chose a cut-off value of log₂FC (<-0.58 and $>+0.58$) the p value (<0.05) when evaluating resistance against bedaquiline and capreomycin.

Thus, even though the log₂FC was set quite low for the initial calculations, we used log₂FC = (<-0.58 and >+0.58), when we eventually labelled an operon as significantly differentially expressed. This equates to a log₂(FC) = 0.58, or a FC of 1.5. That is, under RIF stress, the average expression of the operon must have changed by at least 50%.

2.3 Functional annotation of operon genes

We then carried out a biological process (BP) enrichment analysis to analyze the potential consequences of the operon lengths changes and operon dysregulations under RIF stress.

We did a protein-protein interaction (PPI) analysis and pulled all the functional annotations of the genes belonging to the 50 EVOs from the STRING-DB (<https://string-db.org/>). We also included all genes that were classified among the FP operons to observe if the functions of these genes were predicted as co-expressed neighbouring genes. The Application Programming Interface (API) of the STRING-DB was accessed using a Python script. We also queried Mycobrowser (Kapopoulou et al. 2011), PATRIC (Wattam et al. 2014) and literature to confirm the enrichments, especially if STRING-DB listed the function of the protein as unknown or as a hypothetical protein.

2.4 Testing COSMO on *Mtb* strains under hypoxia

Lastly, we observed whether COSMO was able to capture additional operons when exposed to a different environmental stress. RNA-seq datasets of nine drug sensitive *Mtb* strains (from L4) that were exposed to various levels of hypoxia were obtained from the European Nucleotide Archive (ENA) (<https://www.ebi.ac.uk/ena/browser/home>) under the project code PRJNA478238 (Peterson et al. 2020). The lineage and drug resistance profiles were determined using *TBProfiler* (Phelan et al. 2019) in Galaxy. The fastq RNA-seq reads were aligned to H37Rv (NC_00096.3) using *bwa-mem*, in Galaxy. Subsequent bam files

were then submitted to COSMO under the same cut-offs as the RIF bam files. The hypoxia predicted operons were then analyzed against the EVOs.

3 Results

3.1 Total number of operons

As shown in **bold** (without asterisks) in **Table 7**, 70% of the 50 experimentally verified operons ($n = 35$) were correctly predicted by COSMO across the genotypes. These were predicted by some and not all 64 samples, and some were identified under control conditions and others only under RIF stress.

Five additional operons were predicted for the Family X, WT and *rpoB* mutants, and were not previously predicted when only the Beijing genotype was analyzed.

With the prediction of additional operons, COSMO's performance metrics naturally improved. While the sensitivity increased to **90%** (up from 88%), the greatest improvement was observed in COSMO's precision, which increased from 65% to **76%**. Similarly, the F1 score showed a moderate, but welcomed improvement, from 75% to **82%**.

3.1.2 Variance of samples

As displayed in **Table 6**, the CVs for the BRs, were all lower than the mean (less than 100%), with the average CoV at 20%. For the TR, this average was even lower (12%) – data not shown. The samples with the most TRs ($n = 5$), were not always the most volatile. A sample from the Family X high MIC **experimental** group had five TRs but also had the lowest CoV (8.6%). Nevertheless, the CVs across all the genotypes were all below 100%. Thus, the predictions made by COSMO were reproducible between BRs and TRs. All data and calculations for both TRs and BRs can be found in our Supplementary data on our GitHub page at:

<https://github.com/SANBI->

[SA/COSMO/blob/master/Supplementary_data/CoVs.xlsx](https://github.com/SANBI-SA/COSMO/blob/master/Supplementary_data/CoVs.xlsx)

Table 6: The mean, standard deviation and coefficient of variance of COSMO's predictions for samples of the same genotype

Operons	Number of biological samples with the same prediction call										rpoB mutants	
	WT	Beijing (Low)		Beijing		Family X (Low)		Family X (High)		rpoB mutants		
	CONT	CONT	EXP	CONT	EXP	CONT	EXP	CONT	EXP	CONT	EXP	
1 Rv0046c - Rv0047c	3	3	3	3	3	3	2	3	2	3	3	
2 Rv0096 - Rv0102	2	3	2	3	2	2	2	3	3	2	2	
3 Rv0166 - Rv0178	2	2	2	3	3	3	2	2	2	3	2	
4 Rv0287 - Rv0288	2	2	3	2	3	1	3	1	2	2	3	
5 Rv0490 - Rv0491	3	3	3	3	3	3	3	3	3	3	3	
6 Rv0586 - Rv0594	3	3	3	3	3	1	2	1	2	3	2	
11 Rv0676c - Rv0677c	3	3	3	3	3	3	3	3	3	3	3	
12 Rv0735 - Rv0736	2	3	2	3	2	1	2	1	1	2	2	
13 Rv0902c - Rv0903c	3	3	3	3	3	3	3	3	3	3	3	
14 Rv0928 - Rv0930	3	3	3	3	3	3	3	3	3	3	3	
15 Rv0933 - Rv0936	3	3	3	3	3	3	3	3	3	3	3	
16 Rv0967 - Rv0970	3	3	3	3	3	3	3	3	3	3	3	
17 Rv0986 - Rv0988	3	3	3	3	3	3	3	3	3	3	3	
18 Rv1138c-Rv1139c	3	3	3	3	3	3	2	1	1	3	2	
19 Rv1161 - Rv1164	3	3	3	3	3	3	3	3	3	3	3	
20 Rv1285 - Rv1286	3	3	3	3	3	2	2	1	1	2	2	
21 Rv1303 - Rv1312	3	3	3	3	3	3	3	3	3	3	3	
22 Rv1334 - Rv1336	3	3	3	3	3	3	3	3	3	3	3	
23 Rv1410c - Rv1411c	3	3	3	3	3	3	3	3	3	3	3	
24 Rv1460 - Rv1466	3	3	3	3	3	3	3	3	3	3	3	
25 Rv1477-Rv1478	3	3	2	3	3	3	3	3	3	3	2	
26 Rv1483-Rv1484	3	3	3	2	2	3	3	3	3	3	3	
27 Rv1660 - Rv1661	3	3	3	3	3	2	2	2	1	3	3	
28 Rv1806 - Rv1809	2	3	3	3	3	3	2	2	2	2	2	
29 Rv1826 - Rv1827	3	2	3	3	3	1	2	3	3	3	3	
30 Rv1908c - Rv1909c	3	3	3	3	2	1	2	1	1	3	3	
31 Rv1964 - Rv1966	2	3	2	3	3	1	1	1	1	3	1	
32 Rv1966 - Rv1971	3	3	3	3	3	2	1	1	1	2	3	
33 Rv2243 - Rv2247	3	3	3	3	3	3	3	3	3	3	3	
34 Rv2358 - Rv2359	3	3	3	3	3	3	3	3	3	3	3	
35 Rv2430c - Rv2431c	3	3	3	3	3	3	3	3	3	3	3	
36 Rv2481c - Rv2484c	3	2	3	3	3	2	2	2	3	2	2	
37 Rv2592c - Rv2594c	3	3	3	3	3	3	3	3	3	3	3	
38 Rv2686c - Rv2688c	3	3	3	3	2	3	3	3	3	3	3	
39 Rv2743c - Rv2745c	2	2	2	1	2	1	1	1	1	3	3	
40 Rv2871 - Rv2875	3	3	3	3	3	3	3	3	3	3	3	
41 Rv2877c - Rv2878c	3	3	3	3	2	2	3	2	2	3	3	
42 Rv2931 - Rv2938	3	3	3	3	3	3	3	3	2	3	3	
43 Rv2958c - Rv2959c	3	3	3	3	3	3	3	3	3	3	3	
44 Rv3083 - Rv3089	2	2	2	2	2	2	2	1	1	2	2	
45 Rv3132c - Rv3134c	3	3	3	3	3	3	3	3	3	3	3	
46 Rv3145 - Rv3158	3	3	3	3	3	3	3	3	3	3	3	
47 Rv3417c - Rv3423c	3	3	3	3	3	2	2	3	3	3	3	
48 Rv3493c - Rv3501c	3	2	2	3	2	2	1	3	3	3	3	
49 Rv3516 - Rv3517	3	2	2	3	3	2	3	1	1	2	3	
50 Rv3612c-Rv3616c	3	3	3	3	2	3	2	1	2	3	3	
51 Rv3793 - Rv3795	3	3	3	3	3	3	3	3	3	3	3	
52 Rv3874 - Rv3875	3	3	3	3	3	3	3	3	3	3	3	
53 Rv3917c - Rv3919c	2	2	2	2	2	2	2	2	3	3	3	
54 Rv3921c - Rv3924c	3	3	3	3	3	3	3	3	3	3	3	
No of samples	3	3	3	3	3	9	11	10	13	3	3	
Sum	141	142	139	144	139	126	126	120	122	141	138	
Standard deviation	0.384	0.367	0.414	0.382	0.414	0.728	0.64	0.849	0.804	0.384	0.472	
Mean	2.82	2.84	2.78	2.88	2.78	2.52	2.52	2.4	2.44	2.82	2.76	
CV	13.62	12.91	14.9	13.25	14.9	28.88	25.397	35.36	32.95	13.62	17.09	

3.2 Static operon calls

Table 7 shows that 46% of the 50 EVOs ($n = 23$) were always the same length across the 64 samples, regardless of whether they were grown under control conditions or under RIF-stress. We called these static operons. For 16 of these 23 static operons, all 64 samples matched the operon length of the EVOs. Since they were therefore TPs, and they were static, we called them static TPs. Four of the static operons were always longer than the EVOs for all 64 samples (static FPs) under both control condition and RIF stress, and three operons were consistently shorter across all 64 samples (static FNs).

Table 7: The correctly identified operons and the operons which were static across genotypes.

EVOs	CONT or EXP	CONT & EXP (TP)	CONT & EXP (FP)	CONT & EXP (FN)
Rv0046c-Rv0047c	✓	×	×	×
Rv0096-Rv0102	✓	×	×	×
Rv0166-Rv0178	✓	×	×	×
Rv0287-Rv0288	✓	×	×	×
Rv0490-Rv0491	✓	✓	×	×
Rv0586-Rv0594	✓	×	×	×
Rv0676c-Rv0677c	✓	✓	×	×
Rv0735-Rv0736	✓	×	×	×
Rv0902c-Rv0903c	✓	✓	×	×
Rv0928-Rv0930	✓	✓	×	×
Rv0933-Rv0936	✓	✓	×	×
Rv0967-Rv0970	✓	✓	×	×
Rv0986-Rv0988	✓	✓	×	×
Rv1138c-Rv1139c	✓	×	×	×
Rv1161-Rv1164	×	×	×	×
Rv1285-Rv1286	×	×	×	×
Rv1303-Rv1312	✓	✓		
Rv1334-Rv1336	×	×	✓	✓
Rv1410c-Rv1411c	✓	✓	×	×
Rv1460-Rv1466	✓	✓	×	×
Rv1477-Rv1478	✓	×	×	×

Table 8: The correctly identified operons and the operons which were static across genotypes (*continued*).

Rv1483-Rv1484	×	×	×	×
Rv1660-Rv1661	×	×	×	×
Rv1806-Rv1809	✓	×	×	×
Rv1826-Rv1827	✓	×	×	×
Rv1908c-Rv1909c	✓	×	×	×
Rv1964-Rv1966	✓	×	×	×
Rv1966-Rv1971	✓	×	×	×
Rv2243-Rv2247	×	×	×	✓
Rv2358-Rv2359	✓	✓	×	×
Rv2430c-Rv2431c	✓	✓	×	×
Rv2481c-Rv2484c	×	×	×	×
Rv2592c-Rv2594c	✓	✓	×	×
Rv2686c-Rv2688c	✓	×	×	×
Rv2743c-Rv2745c	✓	×	×	×
Rv2871-Rv2875	×	×	×	✓
Rv2877c-Rv2878c	×	×	×	×
Rv2931-Rv2938	×	×	×	×
Rv2958c-Rv2959c	×	×	✓	×
Rv3083-Rv3089	✓	×	×	×
Rv3132c-Rv3134c	✓	✓	×	×
Rv3145-Rv3158*	✓	✓	×	×
Rv3417c-Rv3423c	✓	×	×	×
Rv3493c-Rv3501c	×	×	×	×
Rv3516-Rv3517	✓	×	×	×
Rv3612c-Rv3616c	✓	×	×	×
Rv3793-Rv3795	×	×	✓	×
Rv3874-Rv3875*	✓	✓	×	×
Rv3917c-Rv3919c	×	×	✓	×
Rv3921c-Rv3924c	×	×	×	×

✓ operon was predicted.

× operon was not predicted.

3.2.1 Functional Annotation (FA) of Static TPs

We further investigated these 16 static TP operons, to have a better understanding of the biological functions that *Mtb* seemed to exert under selection pressure. Again, we know that these are predictions, but if COSMO always called these the same regardless of genotype or experimental conditions, then the expression levels of these genes, either remained the same, or the expression levels of the adjacent genes always changed with respect to one another, suggesting co-expression/co-regulation.

We submitted the 65 genes belonging to these static operons to STRING-DB. STRING-DB shows the PPI and the gene ontology (GO) enrichment against a random selection of proteins from the *Mtb* genome. The random selection of the proteins is carried out automatically by the STRING-DB algorithm. On the highest confidence level, all 65 proteins could be grouped under just three biological processes (BPs). The proteins highlighted in red in Figure 9 show that many proteins were involved in transmembrane transport – more specifically, the transport of various molecules/protons/electrons related to adenosine triphosphate (ATP) synthesis (FDRs $\leq 1.9 \times 10^{-12}$). The blue and yellow-coloured proteins were involved in ‘response to stimulus’ and to ‘pathogenesis’ (FDRs: 0.00014 and 0.0003), respectively. The list of 65 TP operon genes can be found in Supplementary Table 1.

To add to the PPI analysis, we also interrogated the literature, to uncover what others may have found about these operons. We also retrieved the BP for the individual genes of the operons independently using Mycobrowser and PATRIC.

Table 9 shows the most enriched BPs to which the static TP operons were assigned. The bottom of the table provides a summary of the number of operons and their proteins which were associated with a BP, to show that sometimes a BP can be

enriched, due to a large number of proteins linked to it. However, from an operon view, the same BP may not be as enriched. In order of the greatest to the least number of operons targeted, the categories were: i) **cell and cell wall processes** with

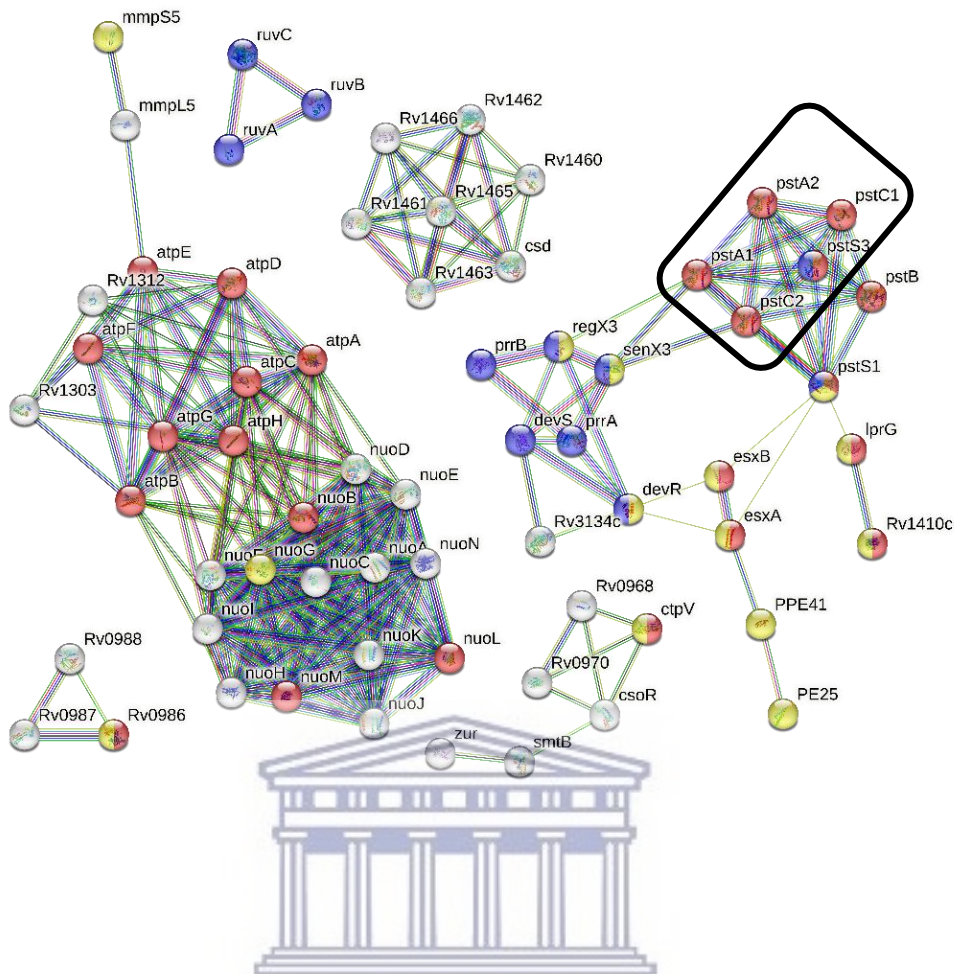


Figure 9: The 65 genes belonging to the static TP operons were submitted to STRING-DB for a functional annotation analysis. The red proteins were linked to transporting of molecules/protons/electrons involved in ATP-related processes (FDR 1.9×10^{-12}). The blue and yellow genes were involved in 'response to stimulus' and to 'pathogenesis' (FDRs: 0.00014 and 0.0003), respectively.

20 proteins across 9 operons, ii) **regulatory proteins** with 10 proteins across 6 operons and iii) **intermediary metabolism and respiration** with 28 proteins across 3 operons. The rest of the 7 proteins were scattered across four operons and belonged to the BPs: 'conserved hypothetical proteins', 'virulence, detoxification, adaptation', 'information pathways' and 'PE/PPE family'.

We then examined whether our proteins/operons may have been allocated to these specific BPs merely because of their prevalence in the genome. (Lamichhane 2018) showed that the most common biological processes to which *Mtb* H37Rv proteins were allocated, were indeed 'i) conserved hypothetical proteins', ii) '**intermediary metabolism and respiration**' and iii) '**cell wall and cell processes.**' The latter two categories for our static TP proteins could therefore have been enriched BPs by chance.

Similarly, the ATP-related processes highlighted by the STRING-DB analysis, is a subcategory of respiration. The reason for this enrichment by STRING-DB was due to 3 operons which each contained a high number of proteins involved in intermediary metabolism and respiration, namely: Rv1303 - Rv1312 (9 proteins), Rv1460 - Rv1466 (5 proteins) and Rv3145 - Rv3158 (14 proteins). That is, 43% of the 66 static TP proteins were involved in ATP-related processes. However, this does not negate the fact that there seems to be a preference for *Mtb* to regulate ATP-related proteins in the form of operons and more specifically, in operons which are under selection pressure. This is BP is also worth noting since the 'conserved hypothetical proteins', which is the most abundant BP for H37Rv, seemed to have been **preferentially avoided** for constituents of static TP operons. In addition, other studies have shown that intracellular concentrations of ATP and proteins involved in the electron transport chain are unusually tightly regulated in *Mtb* to reduce its sensitivity to drugs (Rao et al. 2019; Talwar et al. 2020).

Even more intriguing was that **none** of the static TP operons contained proteins which partook in '**lipid metabolism**', despite this also being one of the biggest functional classes to which H37Rv were assigned. The '**regulatory proteins**' was another class which piqued our interest. Although it was not one of the most prevalent functional categories for H37Rv, they were one of the most enriched BPs for static TP operons. It is therefore possible that, just as with ATP-related proteins, regulatory proteins may be preferentially grouped into operons by *Mtb*.

Table 9: The most common BPs for the 16 static TPs operons across the genotypes.

OPERON	BIOLOGICAL PROCESS OF OPERON GENES						
	RP ¹	CCPW ²	CHP ³	IMR ⁴	PE/ PPE ⁵	IP ⁶	VDA ⁷
Rv0490 - Rv0491	✓	×	×	×	×	×	×
Rv0676c - Rv0677c	×	✓	×	×	×	×	×
Rv0902c - Rv0903c	✓	×	×	×	×	×	×
Rv0928 - Rv0930	×	✓	×	×	×	×	×
Rv0933 - Rv0936	×	✓	×	×	×	×	×
Rv0967 – Rv0970	✓	✓	✓	×	×	×	×
Rv0986 – Rv0988	×	✓	×	×	×	×	×
Rv1303 - Rv1312	×	✓	×	✓	×	×	×
Rv1410c-Rv1411c	×	✓	×	×	×	×	×
Rv1460-Rv1466	✓	✓	×	✓	×	×	×
Rv2358 – Rv2359	✓	×	×	×	×	×	×
Rv2430c – Rv2431c	×	×	×	×	✓	×	×
Rv2594c - Rv2592c	×	×	×	×	×	✓	×
Rv3134c-Rv3132c	✓	×	×	×	×	×	✓
Rv3145 – Rv3158	×	×	×	✓	×	×	×
Rv3874 – Rv3875	×	✓	×	×	×	×	×
Number of Operons	6	9	1	3	1	1	1
Number of Proteins	10	21	1	28	2	3	1

¹Regulatory protein

² Cell and cell wall processes

³ conserved hypothetical protein

⁴intermediary metabolism and respiration

⁵ PE/PPE family proteins

⁶ information pathways

⁷ virulence, detoxification, adaptation



operon contains genes which function in the BP listed in the column header.



operon does not contain genes which function in the BP listed in the column header.

3.2.2 Functional annotation of Static False positives

We then analyzed the four operons predicted in the static FP group, to observe if there was a functional relevance to COSMO adding these additional genes, or if it was just as a result of the crude logic behind the algorithm.

The four-gene operon, Rv1334 – Rv1336, was consistently predicted as a 10 gene operon (Rv1332 – Rv1341) across all genotypes. As shown by the **red** proteins in **Figure 10A**, all of the predicted FP proteins, except for Rv1338 (*murI*), form part of the molybdopterin cofactor (MoCo) metabolic process (FDR = 9.34×10^{-8}) – which is the same metabolic process to which three of the operon's proteins belong. These three EVO proteins also function in a novel cysteine biosynthesis pathway (**blue** proteins). Interestingly, the MoCo metabolic pathway was proven to be tightly linked to the cysteine pathway (Voss et al. 2011; Mendel 2013; Leimkühler 2014).

Operon Rv3793 – Rv3795 was predicted as a range of different lengths for each genotype, but it was always a combination of the proteins from Rv3788 – Rv3798. STRING-DB showed that four of these FP proteins (Rv3789 – Rv3792) which are directly upstream of the EVO, function in the same biological process as the EVO, namely **cell wall organization** (FDR = 1.60×10^{-9}). This is depicted by the **red** proteins in **Figure 10B**. Interestingly, in 2008 Goude et al. carried out an *Mtb* experiment involving ethambutol and ofloxacin treatment and showed that this operon should officially be extended upstream to include three of these FP proteins which COSMO predicted, namely: Rv3790 - Rv3791 (involved in **lipid metabolism and cell wall organization**) and Rv3792 (involved in **cell wall organization**). Hence,

four of the FP proteins take part in the same BP as the EVO and **three** of the FP proteins were confirmed to be co-expressed with this operon when *Mtb* is grown under other drug stresses. It is therefore not improbable that this operon may be extended by some of these proteins, but an appropriate study with the required environmental stress, has not yet been designed to observe their co-expression. It's also interesting that a lipid metabolism protein was implicated.

The Rv2958c - Rv2959c operon was always predicted as being extended by the hypothetical protein Rv2960c. The proteins of the EVO participate in **intermediary metabolism and respiration**. However, the function of Rv2960c is unknown, so it is unclear whether it could be a protein that is occasionally co-expressed with this operon.

Finally, the proteins of operon Rv3921c – Rv3924c, were classified as belonging to a mixture of three BPs, namely: '**cell wall and cell processes**', 'virulence, detoxification, adaptation' and 'information pathways.' This operon was always predicted to overlap with another operon Rv3917c – Rv3919c. These two operons are separated from each other by just one hypothetical protein, Rv3920c. Perhaps it's just proximity, but we may not dismiss that it may be a longer operon, for which previous studies have not recreated the appropriate conditions for its induction.

For a more detailed overview of the functional annotations of these genes, see **Supplementary Table 2**.

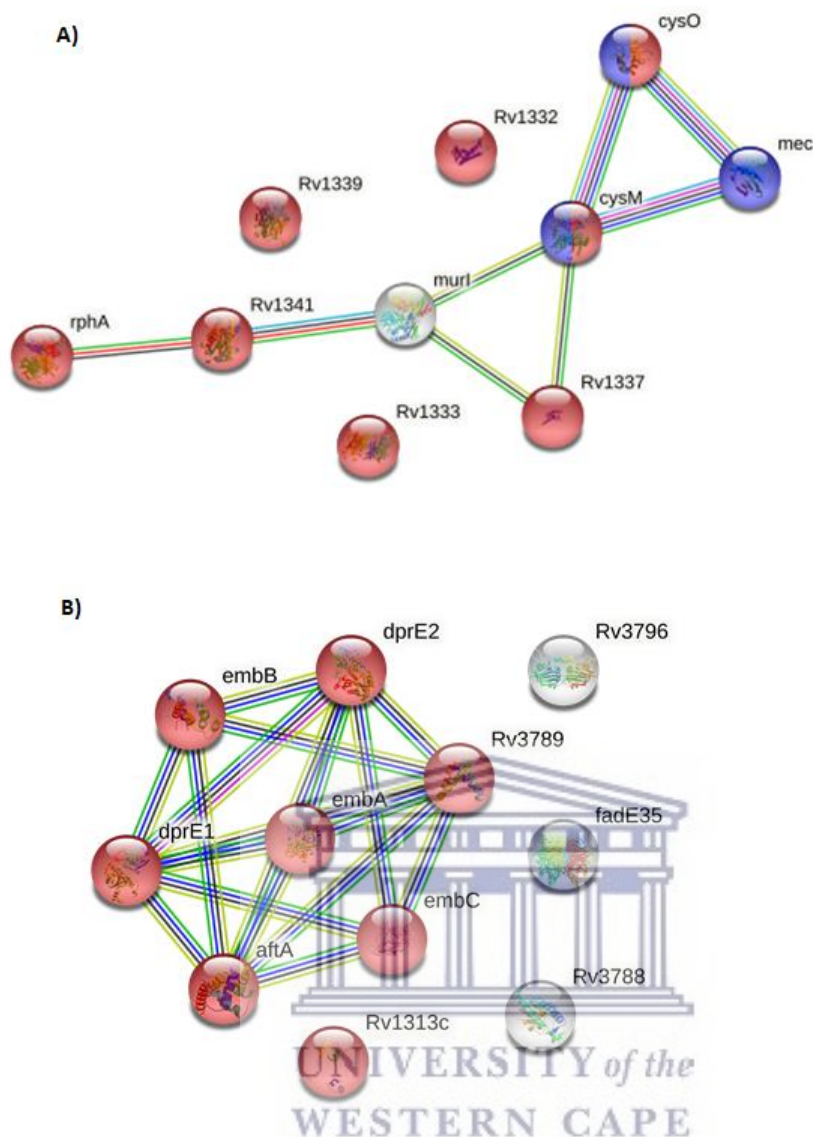


Figure 10: STRING-DB PPI of true positive operons and their predicted false positive genes. A) All of the predicted FP proteins of the operon Rv1334 – Rv1336, except for Rv1338 (*murl*), form part of the molybdopterin cofactor (MoCo) metabolic process (FDR = 9.34×10^{-8}), including three of the proteins which belong to the EVO. The proteins belonging to the experimentally verified operon (EVO) also function in a novel cysteine biosynthesis pathway. However, the MoCo metabolic pathway has been shown to be closely associated with the cysteine pathway. **B)** The genes of the EVO, Rv3793 – Rv3795, and its four upstream FP genes, Rv3789 – Rv3792, have been shown to function in the same BP, namely cell wall organization (FDR = 1.60×10^{-9}). Three of the FP proteins (Rv3790 – Rv3792) have been shown to form part of the operon under two other drug stresses. These are the proteins shown in **red**. For the official gene names of these common gene names, see **Supplementary Table 2**.

3.2.3 Functional annotation of Static False Negatives

Finally, we assessed the possible reasons for why some operons were consistently predicted as shorter than the EVOs. Operon Rv1161-Rv1164 was always split, with the first two genes expressed as single genes, while the latter two were expressed as an operon. The reason behind the split was a huge drop in the IGR expression levels between these genes (up to a 14x fold difference). Since the IGRs and their flanking CDSs usually showed correlated expression in experimental operon studies (REFs), it's unlikely that these genes were coregulated in our strains. All of these genes are involved in **intermediary metabolism and respiration**.

For the *kas* operon (Rv2243 - Rv2247), the first gene of this operon (*fabD*), was always expressed independently from the rest of the operon. This was as a result of a 3-fold higher expression level compared to the rest of the genes. Consistent with our findings, both Wilson et al. (1999) and Fu (2006) showed that *fabD* showed varying expression levels, depending on which drug it was exposed to. Under isoniazid exposure, it was **increased by** up to 3-fold compared to the rest of the operon genes, while a 2-fold **decrease** was observed under delamanid treatment. Moreover, Salina et al. (2019) showed that while all five of the *kas* operon genes were upregulated during an early resuscitation growth stage, the *fabD* was not as highly upregulated as the other four genes. This demonstrates again that different environments induced different lengths in this operon. We found no evidence in literature of this operon being tested under RIF stress. Yet, based on the observations of its modifications in response to other drugs, it's possible that experimental evidence may confirm our results. All the genes of this operon function in **lipid metabolism**.

Lastly, across the different genotypes and conditions, the operon Rv2871 - Rv2875 was never consistently the same length. Rv2871 (*vapB43*) is a possible **antitoxin**, while Rv2872 (*vapC43*) may be a **toxin**, which is involved in antibiotic stress response, **cell wall** structure, and biofilm development (Wang et al. 2018). **Lipid -**

especially cholesterol-induced antibiotic persistence, have been shown to be critically dependent on these toxin-antitoxin systems (Talwar et al. 2020).

Rv2873 (*mpt83*) may be an antigen in the presence of the drug vancomycin (Mustafa 2011) and Rv2875 (*mpt70*) is an immunogenic protein. Both are possibly involved in **cell and cell wall processes**. Rv2874 (*dipZ*) is possibly involved in **intermediary metabolism and respiration**. However, not much is reported in literature about this operon and the functions of most of the proteins are still probable. Hence, we could not do any further cross referencing to experimental studies.

Interestingly, while lipid metabolism proteins were never part of the **static TP** operons, both the static FN and static FP operons had operons lipid-related proteins. This suggest that proteins involved in lipid-related BPs may be a target for adjusting operon lengths.

For a more detailed overview of the FA of these genes, see **Supplementary Table 2**.

3.3 Dynamic operons

Unlike the 23 static operons where the operon lengths were the same across all 64 samples, for **27** of the 50 EVOs (54%), the operons were **not** consistently the same length across samples. Instead, the strains belonging to each genotype reorganized the operons in their unique way under control conditions or RIF stress. Where some strains may have lengthened an operon in response to RIF stress (FP), other strains may have shortened the same operon (FN), and still others may have retained the operon length, even after being exposed to RIF stress. We therefore called these **dynamic operons**, because the operon had different lengths amongst the strains.

Table 10 shows that although in response to RIF stress, the strains regulated their operon lengths uniquely, there was still a general sense of resistance to operon length modifications. Moreover, if the modifications were allowed, the type of operon length modification seemed to be similar amongst strains of the same genotype. That is, on average, for 63% of these dynamic operons ($n = 17/27$), the strains of a genotype almost agreed upon what the length of the operon should be for their specific genotype under control conditions and would resist a change in length when RIF stress was introduced.

***Note that this was 63% of dynamic operons ($n = 27$), and not of the 50 total EVOs.**

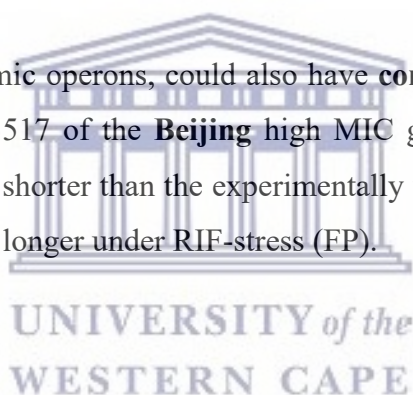
3.3.1 Heterogeneous calls versus consensus calls for dynamic operons

A distinction was made between a dynamic operon and an operon prediction call that was a **consensus call** or a **heterogeneous call**. Just to reiterate, a dynamic operon refers to an operon which is not the same length across all 64 samples under control **AND** RIF-stress conditions. A consensus prediction by COMSO was one where all biological replicates of a genotype agreed on the prediction call under a growth condition. For example, all strains of a genotype agreed that their operon was longer than the experimentally verified operon (FPs) under either control conditions or RIF stress. On the contrary, a heterogeneous call, was one where at least 1 of the biological replicates (BRs), received a prediction call that was different from the rest, under either control conditions or RIF-stress.

Dynamic operons could therefore have a consensus **or** heterogeneous prediction call. Static operons could have **ONLY** consensus calls, because for these all 64 samples had the same prediction call, which is why the topic of heterogeneity was never introduced then.

Table 10 illustrates the prediction outcomes for dynamic operons. One example was that of operon Rv0735 – Rv0736 from the Beijing high MIC genotype, for which COSMO predicted the operon as TP under control conditions, but under RIF-stress, some operons remained TP, while others were predicted as longer (FP), resulting in a **heterogeneous call**, with a combination of TPs + FPs operons. An alternative scenario for a heterogeneous call, was when COSMO identified a complete mixture of operon lengths, such as with operon Rv2743 – Rv2745, where for the same genotype, this operon received all three prediction calls (FP, FN and TP) under control conditions.

On the contrary, dynamic operons, could also have **consensus calls**, such as with operon Rv3516 – Rv3517 of the **Beijing** high MIC genotype, where all strains reported the operon as shorter than the experimentally verified operon (FN) under control conditions, but longer under RIF-stress (FP).



3.3.2 Operon change genotype or strain-specific?

One of the primary questions we had was how often the operons within a genotype had a consensus call, and how often the operon had a heterogeneous call. This was important because the former could allude to a genotype-specific regulator/promoter which modulates the length of an operon for that genotype, while the latter means that each strain regulates its operon length independently. We also wanted to know whether this possible genotype-specific regulation occurred more often under control conditions or under RIF-stress.

Table 10 shows that for the 27 dynamic operons, although strains were slightly more likely to have heterogeneous calls under RIF stress ($p = 0.02$), under both control conditions and RIF-stress conditions, strains belonging to the same genotype were more likely to have consensus calls than heterogeneous calls ($p = 0.0006$ and $p = 0.01$, respectively). This therefore still indicates that strains of the same genotype tended to modify operons as if they were regulating their gene expression via operons according to their genotype, and not as independent strains.

Additionally, when we consider that our strains already shared 23 static operons for which the lengths of the operons remained the same across the 64 samples, and we combine this data with the **consensus calls** of the dynamic operons (average $n = 17$), then out of this **combined** EVO operons, approximately 80% ($n = 40$) of strains had consensus calls for their operon lengths ($p = 1.4 \times 10^{-9}$). This indicates that, as a whole, *Mtb* strains showed a greater propensity for regulating their operon lengths in a genotype-specific manner than for each strain to respond to environmental stresses autonomously.

We also checked if certain genotypes showed a greater affinity for operon length modifications under RIF stress than others. The **low** MIC **Family X** strains from L4 more readily modified their operon lengths under RIF-stress compared to the other genotypes ($n = 16$; **32%**), while the Beijing low MIC strains offered the greatest resistance to length changes ($n = 6$, 12%). Nevertheless, despite being of the same family, the **high** MIC strains of **Family X**, contrarily reported some of the greatest **resistance** to being modified ($n = 8$; 16%). Hence, it was not just the family, but also the MIC status of strains which influenced how operon lengths were adjusted in response to RIF exposure.

Table 10: The dynamic operons and their predictions under control conditions versus RIF stress, per genotype.

Dynamic Operons	Beijing (Low MIC)		Beijing (High MIC)		Family X (Low MIC)		Family X (High MIC)		rpOB mutants	
	Lineage 2		Lineage 2		Lineage 4		Lineage 4		Lineage 4	
	CONT	RIF	CONT	RIF	CONT	RIF	CONT	RIF	CONT	RIF
Rv0046c - Rv0047c	FN		FN		FN	TP, FN	FN		FN	
Rv0096 - Rv0102	TP	TP, FN	FN	TP, FN	FP, TP, FN	FN	FN		TP, FN	
Rv0166 - Rv0178	TP, FN		FN		TP	TP, FN	TP, FN		TP	TP, FN
Rv0287 - Rv0288	FP, TP	FP	FP, TP	FP	FP, TP	FP	FP, TP		FP, TP	TP
Rv0586 - Rv0594	TP		TP		FN	TP, FN	FN TP, FN		TP	
Rv0735 - Rv0736	FP	FP, TP	TP	FP, TP	FP, TP	FP	FP, FN, TP	FP, TP	FP, TP	TP, FN
Rv1138c - Rv1139c	FP		FP, TP	FP	FP		FP, TP		FP	FP, TP
Rv1285 - Rv1286	FP		FP		FP		FP, FN	FP, FN	FP, FN	
Rv1477 - Rv1478	FP		FP, TP	FP	FP		FP		FP	FP, TP
Rv1483 - Rv1484	FP		FP, FN		FP		FP		FP	
Rv1660 - Rv1661	FP		FP		FN		FN		FP	
Rv1806 - Rv1809	FN		FN		TP	TP, FN	TP, FN	FN	TP, FN	
Rv1826 - Rv1827	FP, TP	FP	FP	FP	FP, TP		FP		FP	
Rv1908c - Rv1909c	FP		FP	FP, FN	FP	FP, FN	FP		FP	
Rv1964 - Rv1966	FP	FP, FN	FN		FP, FN, TP	FP, FN	FP	FP, FN	FP	FP, FN, TP
Rv1966 - Rv1971	FN		FN		FP, FN	FN	FN	FP, FN	FP, FN	FN
Rv2481c - Rv2484c	FP	FP, FN	FN	FP	FP		FP		FP, FN	
Rv2686c - Rv2688c	FP		FP	FP, TP	TP		TP		FP	
Rv2743c - Rv2745c	FP, FN	FP, FN, TP	FP, TP	FP, TP	FP	FP, TP	FP, FN, TP	FP, TP	TP	
Rv2877c - Rv2878c	FP		FP	FP, FN	FP		FP, FN		FP	
Rv2931 - Rv2938	FP		FP		FP		FP, FN		FP	
Rv3083 - Rv3089	TP, FN	TP, FN		TP, FN		TP, FN		TP, FN		
Rv3417c - Rv3423c	FN		FN		TP, FN	FN	FN		FN	
Rv3493c - Rv3501c	FP, FN	FP	FP, FN	FP, FN	FP, FN	FP	FP		FP	
Rv3516 - Rv3517	TP, FN	FN	TP	TP, FN	TP	TP	TP	TP, FN	TP, FN	TP
Rv3612c - Rv3616c	FN		FN	FP, TP	FN	FP, TP	FP, FN		FN	
Rv3917c - Rv3919c	FP, FN		FP, FN		FP	FP, FN	FN		FN	

* The **orange blocks** were operons which were the same length under RIF stress as they were under control conditions, within their genotype. The **green blocks** indicate operons which changed their length under RIF stress and indicates that the change was a unanimous call (consensus) under both conditions E.g., FN under control conditions and FP under RIF stress. Alternatively, the **green blocks** may also indicate that the operon length was a unanimous call under one condition (e.g., FN under control) and a slightly mixed call in another (e.g., TP and FN under RIF). Grey blocks indicates that the calls were completely mixed (TP and FP and FN) under one or both conditions.

3.3.3 Functional annotation of dynamic operons

Naturally, after observing the most enriched BPs for the static operons, we wanted to analyze the functional annotations of the dynamic operons, to see if this gives us some hints as to the mechanisms used by *Mtb* to circumvent RIF stress.

The three most enriched BPs for the dynamic operons were i) **cell and cell wall processes**, ii) **intermediary metabolism and respiration** and iii) **regulatory proteins**, which are shown in **Figure 11**. As discussed before, the first two BPs could again have been coincidental, as these were two of the BPs containing the most genes on average for *Mtb* H37Rv. Yet, the regulatory protein class, was again one of the most targeted BPs for these dynamic operons, despite not being one of the most enriched protein classes for H37Rv.

More specifically, operons were most often **extended** with **regulatory proteins**, resulting in a longer operon (FP). Since regulatory proteins were also a target for static TP operons, it may demonstrate a propensity for *Mtb* to package this protein class into operons, or more importantly, for their *impromptu* co-expression with existing operons. The aim of an operon is after all, to **regulate** the genome more efficiently.

In contrast, operons were most frequently **split** at a protein involved in **‘lipid metabolism’**, leading to shorter operons. This emphasised yet again, that while none of the static TP operons harboured genes involved in lipid metabolism, this class of proteins may be important target sites for operon adjustments under RIF stress. For a more detailed overview of the FA for all of the dynamic operon genes, see **Supplementary Table 3**.

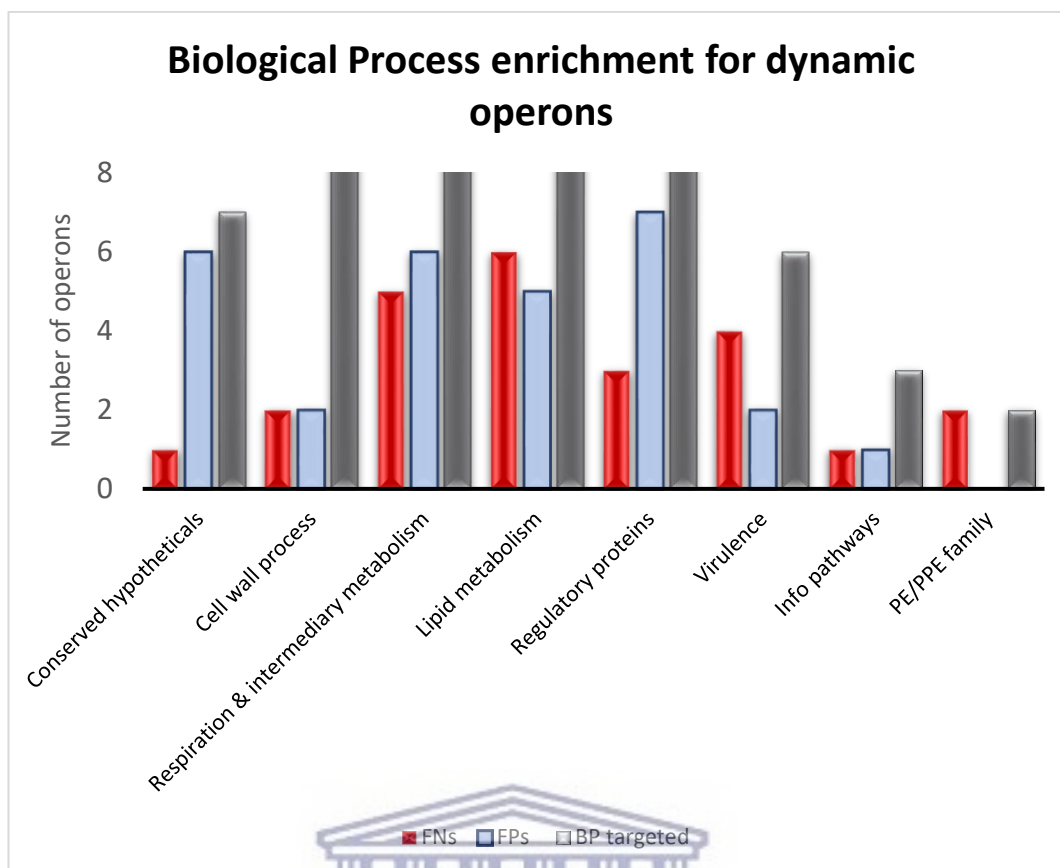


Figure 11: Functional annotation of the genes constituting the dynamic operons. The most common biological processes (BPs) to which the dynamic operon genes belonged were ‘cell and cell wall processes’, ‘intermediary metabolism and respiration’ and ‘**regulatory proteins**’, respectively. Proteins involved in **lipid metabolism** were the most frequent target sites where an operon would be split, leading to FNs. Operons were most often extended by **regulatory proteins**, leading to FPs.

3.3.3.1 Slowed growth rate

In **Chapter 2, Section 2.3.3**, we discussed how operons involved in lipid modifications have a history of being targeted for drug resistance (Derzelle et al. 2004; Bennett and Clarke 2005; Mouammime et al. 2017). A secondary analysis involved a more detailed probing into the specific proteins where the operons were split at lipid metabolism genes and lengthened by regulatory proteins. This revealed that when these specific proteins were differentially expressed or disrupted in previous studies, they always resulted in a **slowed growth rate**, which in turn contributed to the non-replicating survival state of *Mtb* strains.

In some instances, the FN operon resulted from lipid proteins which were downregulated compared to the rest of the operon gene, and previous studies confirmed that those proteins were in fact non-essential (Wipperman et al. 2014). In some cases, they even showed additional internal promoters behind the changes in co-regulation (Bhat et al. 2017).

Lipid homeostasis is tightly controlled by *Mtb* to acclimatize the cell to ever-changing environments and to allow for optimum growth and/or survival (Zhang and Rock 2008), but lipid biosynthesis is an energetically expensive activity (Zhang and Rock 2010). Thus, it makes sense for *Mtb* to inactivate or downregulate these genes unless they are absolutely required for survival.

Conversely, all of the **FP** regulatory proteins by which the dynamic operons were extended, were shown to be essential for persistence under stress and/or they were required for cholesterol uptake, demonstrating that these regulatory proteins were specifically linked to lipid metabolism.

In summary, there were reasonable explanations for the FP or FN prediction calls of these dynamic operons. The different length predictions seemed to point to underlying changes in co-expression when strains were exposed to RIF stress, which was also supported by literature, and which pointed to a careful selection of very specific biological processes to aid in its survival under RIF stress.

For a more detailed look at how these proteins have been implicated in lipid metabolism and slowed growth, see **Supplementary file 1**.

3.4 Differential expression of operons (DEO)

The final part of the analysis was to observe if *Mtb* strains may have opted to up- or downregulate the entire operon's expression, instead of altering the gene expression of individual genes, which may explain the low prevalence of changes in operon lengths under RIF stress. Despite setting the initial threshold very low for the differential expression analysis, most of the operons failed the criteria for evidence of differential expression.

Table 11 shows that while operon Rv0096-Rv0102 had an FDR <0.05, its FC fell just below the cut-off of absolute 0.58 [$|\log_2(FC)| > 0.58$]. Only one operon met both conditions. The *rpoB* mutant strains downregulated operon Rv0676c-Rv0677c by about 60% under RIF stress ($FC = 2^{-0.7}$).

It is not clear why this operon was downregulated. A 2007 study carried out by Briffotiaux et al., reported that this operon is highly conserved in all sequenced genomes of the *Mycobacterium* genus, with the exception of *M. leprae*. This efflux pump, also known as MmpS5-MmpL5, is essential for virulence (Wells et al. 2013) and has been implicated in bedaquiline, clofazimine and azole resistance (Milano et al. 2009; Andries et al. 2014). However, Narang et al. (2019) did show that MmpL5 was only induced in 44% of RIF resistant clinical strains under RIF stress. Nevertheless, despite being downregulated under RIF-stress in the *rpoB* mutants, the expression level of this operon across all genotypes - **including** that of *rpoB*, was typically still quite high (minimum of 50x coverage). Not only was its expression level substantial, but the length of MmpS5-MmpL5 was also previously shown by us to be tightly regulated – even under RIF stress, as this was one of the static TP operons, as reported in **Section 3.2.1**.

Table 11: The differentially expressed operons called by *limma voom*

Genotype	Operon	logFC	Adjusted p-value (FDR)
<i>rpoB</i> mutants	Rv0676c-Rv0677c	-0.7	0.016
<i>rpoB</i> mutants	Rv0096-Rv0102	-0.56	0.016

3.5 Testing COSMO on *Mtb* strains under hypoxia

The last functionality we tested, was COSMO's ability to predict operons under a different environmental stress. By using six drug sensitive strains grown under hypoxia conditions (lineage 4), a combined set of 48% of operons were called. We found that COSMO was able to correctly predict one additional operon (Rv1285-Rv1286), which was never previously picked up under RIF-stress. A STRING-DB analysis revealed that this operon functions in cellular response to sulphur starvation and oxidative stress ($FDR < 2.96 \times 10^{-5}$). Fu and Tai (2009) showed that the first gene of this operon is often upregulated in *Mtb* to survive hypoxia. This finding was supported by Punina et al. (2015) who showed that although these are sulphate adenylyl transferase genes, this operon is also upregulated under hypoxia in *Mtb*.

These results demonstrate that COSMO is able to differentiate between operons induced under different stress conditions (hypoxia versus RIF-stress). Moreover, together with the operons predicted under RIF-stress, this brings the total number of correctly predicted EVOs by COSMO, to 74%. The percentage of true positives is likely to increase when samples grown under other experimental conditions are included in the experimental set.

4 Discussion

After COSMO was benchmarked against Rockhopper, REMap and DOOR 2.0 and was shown to outperform all three algorithms, we wanted to ascertain if operons were up- or downregulated or reorganized when exposed to RIF-stress with respect to their genotype. We tested 64 samples belonging to the WT, *rpoB* mutants, Beijing (high- and low MIC) and Family X (high- and low MIC) genotypes.

We previously showed that although COSMO correctly identified more EVOs than its best performing competitors, it still captured only 60% of full-length operons. However, we are aware that the current understanding is that operons are dynamically responsive to the environment, and thus, they would reorganize depending on which genes are required for survival under that environmental stress. The operons we collected from literature were identified under a wide variety of stresses. Nevertheless, under just one stress (RIF), COSMO was able to predict 70% of full-length operons documented in literature with the addition of the strains from other genotypes. We also showed an improvement in the three metrics used to measure the performance of COSMO, namely, its precision, recall and F1 score. The precision had the greatest improvement (17%).

We discovered that a large percentage of operons (46%) did not modify their length under RIF stress across all the genotypes but maintained the length they had under control conditions (static TP, FP or FN). The biggest group of these static operons belonged to the TP operons (32%). That is, they were the same length as the EVOs, and they never altered their length when exposed to RIF. We investigated their biological functions. They could be collectively grouped under ATP synthesis and transmembrane transport activity related to this process. If these operons were under such intense selective pressure, then they may have to be classified as house-keeping operons. These static TP operons also had an usually high number of regulatory proteins within their operons. Since operons are higher modules that may control entire pathways, targeting them for antitubercular intervention, could prove to be more efficient than individual genes which are the current drug targets.

A small subset of these static operons was always predicted as FP (8%) and FN (6%). Most of the FP proteins were enriched for the same biological process as the EVO which they were ‘falsely’ predicted to be a part of. Interestingly, some of these static FP genes have previously been proposed as extensions to these associated operons, due to the co-expression with the operon genes under certain drug stresses (Goude et al. 2008; Voss et al. 2011; Mendel 2013; Leimkühler 2014).

Similarly, the genes of FN operons were split by COSMO due to unusual discrepancies in expression levels between neighbouring operon genes and/or their uncorrelated IGR coverages. Here again, other studies have confirmed that depending on the environmental stress, many of these genes may not be co-expressed with the rest of the operon genes (Wilson et al. 1999; Fu 2006; Wang et al. 2018; Salina et al. 2019). Some were even shown to specifically have internal promoters to regulate this co-expression (Bhat et al. 2017). Hence, the rules underlying our algorithm are showing promise for its ability to correctly predict operons from literature. A CoV analysis also showed that the predictions of COSMO are reproducible across BRs and TRs.

Regarding strain or genotype-specific cues to stress; we observed that operon lengths among strains were usually the same for strains belonging to the same genotype. Instead of responding to RIF stress autonomously, operons preferred to settle on operon length within their genotype and may be under such intense pressure to maintain that operon length, that they resist operon reorganization, even under RIF stress. On average 80% of operons maintained the length they had under control conditions after they were exposed to RIF stress, within each genotype.

The Family X low MIC strains from L4 were the most flexible to operon reorganizations. However, this same family reported both the lowest and the highest resistance to operon length changes, depending on the MIC levels. Therefore, both the family and MIC status contributed to the strains’ response to RIF stress.

Although this lack of substantial differences in the operon reorganization under RIF stress was a bit disappointing, this resistance to length modifications was supported by literature. Price et al. (2006) observed that there was a strong negative selection against operon reorganizations. Interestingly, Yoon et al. (2011) showed that microorganisms that were resistant to high heat, may also resist changes in operon organization. Although *Mtb* is a mesophile, previous studies have shown that it displays remarkably similar behaviour to thermophiles in its ability to survive high temperatures (Zwadyk et al. 1994; Doig et al. 2002).

A functional annotation of the dynamic operons (operons which allowed length adjustments across genotypes) revealed that proteins functioning in lipid metabolism may be the sought-out target sites where operons receive signals to split in response to RIF-stress. Proteins which function in lipid metabolism have already been consistently highlighted as targets for drug resistance in *Mtb* (Bailo et al. 2015; Bah et al. 2020). On the contrary, regulatory proteins were often preferred as targets where operons were extended (FPs). These specific FP regulatory proteins which were predicted by COSMO, were linked to lipid pathways in literature – pointing again to the careful regulation of lipid levels in response to RIF stress.

When *Mtb* goes into a persistence state, it does so in a well orchestrated process, which starts with it recognizing the type of stress in its immediate environment. This is followed by inducing the formation of the lipid-rich macrophages of its host. *Mtb* then adapts itself for the utilization of these host-derived lipids (especially cholesterol) and lyses the macrophages for lipid uptake. However, the lipid uptake requires the complimentary **regulatory proteins** to activate the **lipid metabolism** enzymes. At the same time, *Mtb* then signals for the **inhibition of growth** pathways, since lipid/cholesterol metabolism is bioenergetically expensive. The slowed growth rate results in *Mtb* persistors, which are able to survive under severe stress. One serendipitous consequence of this, is that *Mtb* becomes less susceptible to drugs (Pandey and Sasseti 2008; Miner et al. 2009; Zhang and Rock 2010; Colangeli et al. 2014; Talwar et al. 2020). This cascade of events shows the intricate

link and therefore possible strategy behind *Mtb*'s targeting of lipid metabolism genes and the associated regulatory proteins, to attain RIF resistance.

When we looked at the differential expression of operons under RIF-stress, only the MmpS5/L5 operon was downregulated in the *rpoB* genotype. We could not ascertain the reason for the downregulation, because MmpS5/L5 is an efflux pump for many anti-tubercular drugs. Despite being downregulated, the average expression levels across all genotypes – including *rpoB* mutants- were still quite high, indicating that this pump was always in an induced state. This operon was also one of the operons for which the length was strictly maintained across all genotypes – even under RIF stress. This intense selection pressure aligns with what was observed in literature and lends to the call that this may be an interesting drug target for *Mtb*.

Nevertheless, the general lack of differential expression of operons, as well as the low prevalence of operon length alterations under RIF stress, seems to indicate that *Mtb* tightly regulates the co-expression of the proteins which constitute operons, so that both their lengths and their overall expression levels are preserved, and if they do fluctuate, they do so mostly in a genotype-specific manner and by specifically targeting proteins which assist in lipid metabolism.

Finally, when COSMO was tested on strains grown under hypoxia, we identified an operon which was previously never called correctly under RIF stress. It was therefore not surprising when the genes of this operon were confirmed to be induced in the presence of hypoxia stress in *Mtb*.

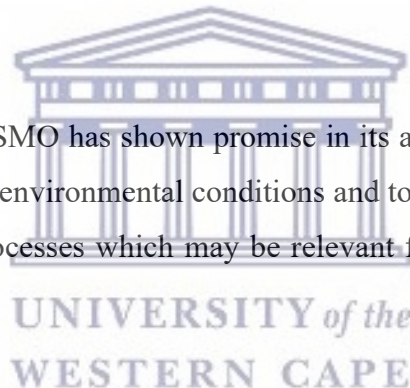
4.1 Limitations of this study

One of the limitations of this study is that only two of *Mtb*'s lineages were analyzed, but this study can naturally be extended to include all *Mtb* lineages, and thus provide a more diverse overview of operon modification. The predicted operons were

also not experimentally validated, so there remains a great scope for increasing the credibility of this study with long range PCR or another appropriate experimental method. Our aim is also for COSMO to accept multiple bam files at once, and thus, COSMO will need to include a means to normalize between samples and possibly even have the capability of processing Fastqs as input.

Furthermore, the EVO used in this study were identified under various conditions – none of which were RIF stress. Also, none of the EVOs were previously compared across different genotypes. There were also no *Mtb* genes which were previously confirmed to always be expressed as single genes. This means that both a true positive and a true negative list were difficult to validate. The lack of a true negative list also prevented us from testing for specificity and from drawing a ROC curve. Lastly, it was a challenge to ascertain how a high GC content and possible gDNA contamination may have influenced the coverage of the various genomic regions under study.

However, overall, COSMO has shown promise in its ability to accurately identify operons under distinct environmental conditions and to give us insight into the underlying biological processes which may be relevant for *Mtb*'s adaptation to RIF stress.



5 Future work

In future we would like to carry out operon predictions for organisms with a more extensive true positive list of operons such as *E. coli* and *B. subtilis*. The current list of EVOs for *Mtb* is very small and therefore creates limitations, especially with regards to the validation of a true negative operon list. We would also like to include samples from other environmental stresses, to observe if more operons from literature can be predicted. We aim to allow COSMO to accept multiple files, which means that it will be optimized to normalize the reads across samples. Future work also includes testing newly predicted operons and dynamically changing operons

in the lab to extend the validated operon list for *Mtb* and includes making COSMO available for non-bioinformaticians on the Galaxy interface.

6 Supplementary Material

Supplementary Table 1: List of 65 genes which made up the static TP operons.

TP genes

Rv0490
Rv0491
Rv0676c
Rv0677c
Rv0902c
Rv0903c
Rv0928
Rv0929
Rv0930
Rv0933
Rv0934
Rv0935
Rv0936
Rv0967
Rv0968
Rv0969
Rv0970
Rv0986
Rv0987
Rv0988
Rv1303
Rv1304
Rv1305
Rv1306
Rv1307
Rv1308
Rv1309
Rv1310
Rv1311
Rv1312
Rv1410c
Rv1411c
Rv1460
Rv1461

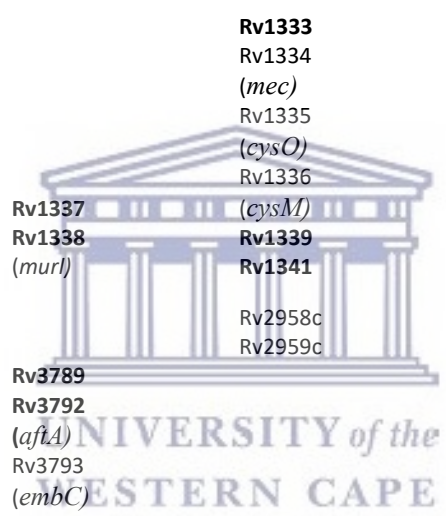


Rv1462
Rv1463
Rv1464
Rv1465
Rv1466
Rv2358
Rv2359
Rv2430c
Rv2431c
Rv2592c
Rv2593c
Rv2594c
Rv3132c
Rv3133c
Rv3134c
Rv3145
Rv3146
Rv3147
Rv3148
Rv3149
Rv3150
Rv3151
Rv3152
Rv3153
Rv3154
Rv3155
Rv3156
Rv3157
Rv3158
Rv3874
Rv3875



Supplementary Table 2: Static FP and static FN operons and the functional annotation of their operon genes

Static FP operons	LM 1	CCPW 2	PE/ PPE 3	IMR 4	VDA 5	IP 6	CHP 7	RP 8	ISP 9
Rv1334 – Rv1336									
Rv2958c - Rv2959c									
Rv3793 - Rv3795									
Rv3921c - Rv3924c									
No. of times BP was targeted									
FPs									
Static FN operons	LM 1	CCPW 2	PE/ PPE 3	IMR 4	VDA 5	IP 6	CHP 7	RP 8	ISP 9



Rv1161 - Rv1164				<u>Rv1161</u> Rv1162 Rv1163 Rv1164					
Rv2243 - Rv2247	<u>Rv2243</u> Rv2244 Rv2245 Rv2246 Rv2247								
Rv2871 - Rv2875		<u>Rv2873</u> <u>Rv2875</u>		<u>Rv2874</u>	<u>Rv2871</u> <u>Rv2872</u>				
No. of times BP was tar- geted		1	1	0	2	1	0	0	0
FNs		1	1	0	2	1	0	0	0

Bold gene name: added protein (FP)

Underlined gene name: protein where the operon was split

*FN: one split was counted between two genes if they were in the same category

¹ Lipid metabolism

² Cell and cell wall processes

³ PE/PPE family proteins

⁴ intermediary metabolism and respiration

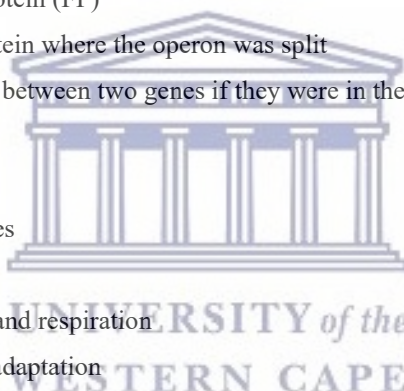
⁵ virulence, detoxification, adaptation

⁶ information pathways

⁷ conserved hypothetical protein

⁸ Regulatory proteins

⁹ Insertion sequences and phages

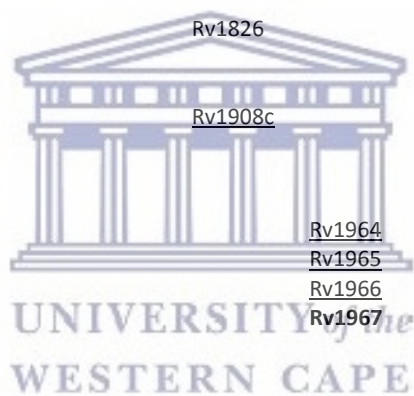


Supplementary Table 3: Dynamic operon genes and the biological process enrichment.

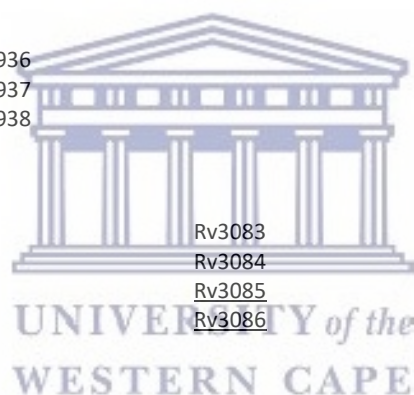
Dynam Opero 1	LM 2	CCPW 3	PE/ PPE 4	IMR 5	VDA 6	IP 7	CHP 8	RP 9
Rv0046c - Rv0047c				<u>Rv0046c</u>				<u>Rv0047c</u>
Rv0096 - Rv0102	Rv0098 Rv0099 Rv0100 Rv0101	Rv0102	<u>Rv0096</u>	<u>Rv0097</u>				
Rv0166 - Rv0178	<u>Rv0166</u>	Rv0173 Rv0175 Rv0176 Rv0178			<u>Rv0167</u> Rv0172 Rv0174			
Rv0287 - Rv0288		Rv0287 Rv0288 Rv0289 Rv0290 Rv0292		<u>Rv0291</u>				
Rv0586 - Rv0594		Rv0593			Rv0587 Rv0588 Rv0589 Rv0590 Rv0591 Rv0592 Rv0594			Rv0586
Rv0735 - Rv0736						Rv0735 Rv0736		Rv0737
Rv1138c - Rv1139c		Rv1139c		Rv1138c				Rv1137c
Rv1285 - Rv1286	Rv1288			Rv1284 Rv1285 Rv1286			Rv1289	Rv1287



Rv1477 - Rv1478		Rv1476 Rv1481	Rv1477 Rv1478	Rv1480	Rv1479
Rv1483 - Rv1484	<u>Rv1483</u> <u>Rv1484</u>		Rv1485		
Rv1660 - Rv1661	<u>Rv1660</u> <u>Rv1661</u> Rv1662 Rv1663 Rv1664 Rv1665				Rv1659
Rv1806 - Rv1809		Rv1806 <u>Rv1807</u> <u>Rv1808</u> Rv1809			
Rv1826 - Rv1827 Rv1908c - Rv1909c		Rv1826 <u>Rv1908c</u>		Rv1825 Rv1829	Rv1827 Rv1828
Rv1964 - Rv1966			<u>Rv1964</u> <u>Rv1965</u> <u>Rv1966</u> <u>Rv1967</u>	Rv1907c	<u>Rv1909c</u>
Rv1966 - Rv1971		Rv1970 Rv1972 Rv1973 Rv1974	<u>Rv1966</u> <u>Rv1967</u> Rv1968 Rv1969 Rv1971		Rv1975
Rv2481c - Rv2484c	Rv2482c Rv2483c Rv2484c			Rv2485c	Rv2481c



Rv2686c - Rv2688c Rv2743c - Rv2745c		Rv2686c Rv2688c Rv2687c	Rv2689c		Rv2690c
	<u>Rv2744c</u>	Rv2743c		Rv2742c	<u>Rv2745c</u>
Rv2877c - Rv2878c	Rv2881c	<u>Rv2877c</u> <u>Rv2878c</u>	Rv2879c Rv2880c	Rv2882c	Rv2883c
Rv2931 - Rv2938	Rv2931 <u>Rv2932</u> <u>Rv2933</u> Rv2934 Rv2935 Rv2939	Rv2936 Rv2937 Rv2938			
Rv3083 - Rv3089	Rv3087 Rv3088 Rv3089		Rv3083 Rv3084 Rv3085 <u>Rv3086</u>		
Rv3417c - Rv3423c		Rv3417c Rv3418c	<u>Rv3419c</u> Rv3421c Rv3422c Rv3423c		<u>Rv3420c</u>
Rv3493c - Rv3501c	Rv3492c Rv3493c Rv3495c			Rv3494c <u>Rv3496c</u> <u>Rv3497c</u> Rv3498c Rv3499c Rv3500c Rv3501c	
Rv3516 - Rv3517	<u>Rv3516</u> <u>Rv3517</u>				



Rv3612c - Rv3616c		Rv3614c Rv3615c Rv3616c					<u>Rv3612c</u> <u>Rv3613c</u>	
Rv3917c - Rv3919c	Rv3916c	Rv3917c <u>Rv3919c</u> <u>Rv3918c</u>						
FNs *	6	2	2	5	4	1	1	3
FPs added	5	2	0	6	2	1	6	7
TP	6	11	1	6	5	1	0	5
No. of times BP was targeted	13	14	2	14	6	3	7	14

Bold gene name: added protein (FP)

Underlined gene name: protein where the operon was split.

*FN: one split was counted between two genes if they were in the same category.

¹ Dynamic operon

² Lipid metabolism

³ Cell and cell wall processes

⁴ PE/PPE family proteins

⁵ intermediary metabolism and respiration

⁶ virulence, detoxification, adaptation

⁸ information pathways

⁸ conserved hypothetical protein

⁹ Regulatory proteins

Supplementary file 1: Literature review of dynamic FP and FN proteins

We probed further into literature to understand the detailed functions of the genes which were target sites where operons were split, to see if they had more in common than just their involvement in lipid synthesis.

We extended this analysis for the genes which were predicted as extensions (FPs) to EVO, to see beyond their enriched BP, 'regulatory proteins.'

Lipid biosynthesis is an energetically expensive process, and genes partaking in these processes have been shown to be silenced during stress to prevent lipids from being oxidized and permanently damaged.

FN under Lipid Metabolism

genes: **Rv0166 - Rv0178**, genotypes: all

The gene involved in lipid metabolism Rv0166 is often downregulated according to COSMO.

When this **gene is disrupted**, *Mtb*'s growth rate slowed down in certain media. However, this disruption gave mice infected with these mutant strains the advantage of surviving 166 days longer.

This operon was often split due to the downregulation of the protein Rv0166 (*fadD5*). Literature confirmed that the **downregulation** of this gene **slows** *Mtb*'s **growth rate**, which enables it to **persist** under stress.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3225055/>

Rv1483 - Rv1484 FN under both conditions for Beijing (High MIC)

This lipid metabolism operon was often split due to uncorrelated expression levels of these proteins and their IGR, which may suggest the use of alternative promoters. This *fabg1* – *inhA* operon is implicated in isoniazid resistance. However, studies have shown that both genes are actually **not required for growth** in *Mtb*.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2393810/>

Rv1660 - Rv1661 split under control and experimental conditions in Family X

When *SigL* was over-expressed, it led to a strong upregulation of four small operons: sigL (Rv0735)-rslA (Rv0736); mpt53 (Rv2878c)-Rv2877; pks10 (Rv1660)-pks7 (Rv1661); and Rv1139c-Rv1138c.

In a murine infection model, the sigL mutant exhibited marked attenuation compared with the parental strain, suggesting a role of σ L in virulence; however, there were no significant differences in the growth rate or in the size and extent of lesions in the infected organs.

<https://febs.onlinelibrary.wiley.com/doi/10.1111/j.1742-4658.2009.07479.x>

[Involved in weakening the cell surface or in an inappropriate modulation of the host immune response https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1418919/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1418919/)

This operon was always split in the Family X genotype, due to the **downregulation** of protein RV1660 (*pks7*). When this specific protein of the operon was **disrupted**, it resulted mutants deficient in lipid biosynthesis and in severe **growth defects** in mice.

<https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.26278-0>

Rv2743c - Rv2745c under control, genotypes: Beijing + Fam X high MIC

Rv2744c may function in regulating lipid droplet homeostasis and nonreplicating persistence (NRP) in *M. tuberculosis*

COSMO showed that Rv2744c (*35kd_ag*) is often significantly **upregulated** compared to the other genes, causing the splitting of the operon. This protein regulates lipid homeostasis and **nonreplicating persistence** (NRP) in *M. tuberculosis*

<https://pubmed.ncbi.nlm.nih.gov/27002134/>

Rv2931 - Rv2938 in Family X high MIC

This operon was often split due to the downregulation of Rv2932 (*ppsB*). Literature has shown that this protein can be specifically targeted for dysregulation to **diminish growth** in *Mtb*.

<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2958.2006.05102.x>

Rv3516 - Rv3517, genotypes: all

This operon was split due to uncorrelated expression IGR between the proteins of this operon, indicating that they may have separate promoters. Others have found that although Rv3516 (*echA19*) is involved in cholesterol metabolism, it is **not essential for growth** when Mycobacteria use cholesterol as fuel during **macrophage** infection.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4255906/>

<https://www.mdpi.com/1420-3049/21/5/598>

When *Mtb* goes into a persistence state, it does so in a well executed process which starts with the recognition of the stress environment. This is followed by the inducing the formation of lipid-rich macrophages. *Mtb* then adapts itself for utilizing host-derived lipids - which includes cholesterol - and lyses its source – the macrophages. The cholesterol uptake then also signals the inhibition of growth pathways, since replicating/growing organisms are susceptible to drugs. At the same time, it signals the activation of cholesterol biosynthesis pathways – resulting in *Mtb* persistors.

<https://dx.plos.org/10.1371/journal.pone.0091024>

<https://pubmed.ncbi.nlm.nih.gov/19634704/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2393810/>

<https://journals.asm.org/doi/10.1128/mSystems.00855-20>



False positives under regulatory proteins

Rv0735 - Rv0736: all genotypes and conditions

This operon was extended by Rv0737. **Rv0737** is a regulatory protein. Some of its functions include adaptation to environmental changes and **resistance** to antibiotics.

<https://www.sciencedirect.com/science/article/pii/S0891584919323767>

Rv1285 - Rv1286 under **RIF** all genotypes and conditions

This operon often included the FP protein **Rv1287** which is a sulfate adenylyl transferase, just like the two proteins of the operon. It uses ATP as its substrate and has been shown to be **upregulated** under nutrient **stress** and implicated in **persistence**.
<https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.2002.02779.x>

Rv1477 - Rv1478 all genotypes

The operon was often extended by Rv1479 (*moxR1*). Hu and Coates (2001) once again, showed that this protein was grossly **upregulated** in *Mtb* **persistor** cells.
<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-6968.2001.tb10780.x>

Rv1660 - Rv1661 FP under control and experimental conditions in the Beijing and *rpoB* genotypes.

This operon was often extended by an upstream protein, Rv1659 (*argH*), and four downstream proteins Rv1662 (*pks8*), Rv1663 (*pks17*), Rv1664 (*pks9*) and Rv1665 (*pks11*). All four downstream proteins have the same function as the two-gene operon and have been shown to be co-expressed with this operon in another study.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1418919/>

This operon was often extended upstream by Rv1659 (*argH*). Argininosuccinate lyase (*argH*) is essential for the **survival** of *Mtb* and plays a key role in nutrient acquisition and **pathogenesis during infection**. It is **required** for growth on **cholesterol**.
<https://onlinelibrary.wiley.com/doi/abs/10.1002/iub.1683>
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5357164/>

Rv1826 - Rv1827 FP in all genotypes

This operon is often extended by Rv1828. The promoter of Rv1828 also lies within the IGR of the operon Rv1826 - Rv1827 and have therefore been suggested to play a role in co-regulating this FP protein and the last gene of the operon. Rv1828 belongs to the *MerR* family of transcriptional repressors/activators, was shown to be **essential** in *Mtb* **infection and survival under stress**.
<https://onlinelibrary.wiley.com/doi/abs/10.1111/febs.14676>

Rv1966 - Rv1971 FP in all except Beijing high MIC

This operon was often extended by protein Rv1975. Despite being a regulatory protein, it is upregulated during **cholesterol** metabolism and is as such **required** for **long-term growth in macrophages**.

<https://europepmc.org/articles/PMC4255906>

Rv2686c - Rv2688c FP in Beijing and *rpoB*

This operon was predicted to be extended by Rv2690c. This transport protein helps to increase apoptosis during **infection of macrophages**.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2864267/>

It was also previously confirmed to be upregulated in *rpoB* mutant strains, helping these strains to increase their uptake of exogenous amino acids.

<https://www.frontiersin.org/article/10.3389/fmicb.2018.02895>

Access to amino acids such as arginine are essential to bacterial infection, but it is a bioenergetically expensive activity. Thus, there is a preference for sequestering these amino acids from the external environment, by inducing the relevant genes.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC127984/>

<https://journals.asm.org/doi/10.1128/JB.00064-09>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1482997/>

Rv2481c - Rv2484c, genotypes: Beijing and *rpoB* (cont and exp)

Rv2481c was a FN regulatory protein, because it was downregulated in some strains of the Beijing and *rpoB* lineage.

Putative triacylglycerol synthase *tgs* genes are induced when the pathogen goes into the non-replicative drug-resistant state caused by slow withdrawal of O₂ and also by NO treatment, which is known to induce dormancy-associated genes.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC451596/>

7 Data Availability Statement

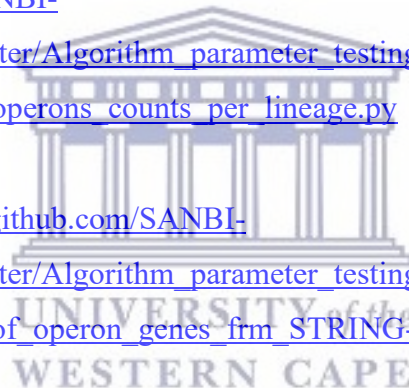
The links to the most important scripts used for this thesis can be found below. All scripts were written by myself.

Operon list: https://github.com/SANBI-SA/COSMO/blob/master/Algorithm_parameter_testing/python_scripts_for_prediction_calls/50_combined_operon_list.txt

Script for making predictions (T, FP, FN): https://github.com/SANBI-SA/COSMO/blob/master/Algorithm_parameter_testing/python_scripts_for_prediction_calls/dict_no_strains_pred_operon_TP_FP_FN.py

Script for summarizing the total predictions across lineages and families:
https://github.com/SANBI-SA/COSMO/blob/master/Algorithm_parameter_testing/python_scripts_for_prediction_calls/predicted_operons_counts_per_lineage.py

STRING-DB: https://github.com/SANBI-SA/COSMO/blob/master/Algorithm_parameter_testing/python_scripts_for_prediction_calls/fetch_FA_of_operon_genes_frm_STRING-db.py



CHAPTER 5

CONCLUSION AND FUTURE WORK

In this thesis we sought to develop a new operon predicting algorithm to detect operons in *Mycobacterium tuberculosis*. We aimed to build COSMO on the foundation laid by REmap, in that it is not constrained by features representative of one organism's genomic architecture. We also aimed to test COSMO across lineages to observe whether operons were reorganized under RIF stress and whether it was done in a lineage-specific or strain-independent way. Finally, we also sought to detect if any operons were targeted for up- or downregulation under RIF stress.

Summary of Chapter 3

In Chapter 3 we developed COSMO by integrating the user-defined parameters: i) the minimum CDS coverage, ii) the minimum IGR coverage, iii) the maximum FD between adjacent CDSs and iv) the maximum FD between CDSs and their flanking IGRs. COSMO also has a built-in feature which re-evaluates the length of the operon upon the addition of each new CDS, by testing whether the averages of all CDS belonging to the operon are still within the FD cut-off. The parameters were empirically validated and found to be statistically significant. We used 12 RIF resistant samples from the Beijing family (L2), which were grown under control conditions and under RIF-stress. COSMO outperformed REmap, Rockhopper and DOOR 2.0 by correctly identifying the most full-length EVOs. COSMO was also able to better distinguish between operons predicted under RIF stress versus those predicted under no RIF stress and obtained the highest F1 score. Our MLRA also showed that the greatest impact on the outcome variable was the new parameter - maximum FD between the IGR and its flanking CDSs and that the least significant parameter was the traditionally used maximum FD between adjacent CDSs.

Summary of Chapter 4

In Chapter 4 we tested COSMO on 64 samples consisting of WT (drug sensitive) strains and RIF resistant strains from Beijing (high- and low MIC), Family X (high- and low MIC) and *rpoB* mutant strains. We later also tested COSMO on nine *Mtb* samples grown under hypoxia that were obtained from literature.

Using COSMO's default settings, we were able to increase the number of operons correctly identified from 60% to 70% of EVOs. We also increased our precision (from 65% to 76%), our recall (from 88% to 90%) and our F1 score (from 75% to 82%).

A larger than expected number of operons (32%) maintained the length of the EVOs, regardless of genotype and even when they were exposed to RIF stress. If these operons are under selective pressure, then these may be valuable higher module targets for anti-tubercular drugs.

We showed that in general *Mtb* tends to resist operon reorganization – even under RIF stress. Approximately 80% of operons had the same call under control conditions as under RIF stress. That is, within a specific genotype ($n = 40$) of sample had consensus calls for their operon lengths ($p = 1.4 \times 10^{-9}$). In the ~20% of cases when operon lengths were modified, the data showed that these strains were more likely to modify their operon lengths within their genotype than to do so independently, under both control ($p = 0.0006$) and RIF stress conditions ($p = 0.01$).

Similarly, only one operon was downregulated under RIF stress - the MmpS5/L5 efflux pump. This pump seems to be important to *Mtb* because it was previously also observed to be one of the operons that was under selective pressure by strictly co-regulating the genes of the operon across all genotypes (static TP). Although it was downregulated, its expression levels were still high under both experimental conditions.

We also showed that ATP-related proteins and regulatory proteins may be preferred constituents of “housekeeping” operons, and the latter were also preferred additions to lengthen operons when *Mtb* was under RIF stress. On the contrary, proteins involved in lipid metabolism seem to be targeted when *Mtb* adjusted its operon lengths in response to RIF stress, resulting in shorter operons.

Finally, under hypoxia, COSMO correctly called one operon that was never previously called under RIF stress. This operon was previously confirmed to be upregulated under hypoxia stress.

Taken together, COSMO successfully demonstrated an improved ability to correctly identify operons, compared to some of the best existing predictors. It is also able to distinguish between operons induced under different stresses and to do this more accurately than its comparators.

Future work

We aim to test COSMO on its ability to predict operons in other bacteria. We also aim to optimize COSMO to accept and process multiple bam files per run. Part of the optimization will involve a model which will allow COSMO to suggest optimal parameters to the user, or to utilize these parameters automatically, based on the expression profiles within the submitted bam files. COSMO will also be tested on samples exposed to alternative stresses and experimental stages to observe if more/different operons are predicted. Finally, for those who are not proficient in Bioinformatics, we aim to integrate COSMO into the Galaxy environment.



UNIVERSITY *of the*
WESTERN CAPE

References

- AHMAD, S., S. EL-SHAZLY, A. S. MUSTAFA and R. AL-ATTIYAH. 2005. 'The Six Mammalian Cell Entry Proteins (Mce3A–F) Encoded by the Mce3 Operon Are Expressed During In Vitro Growth of *Mycobacterium tuberculosis*'. *Scandinavian Journal of Immunology* 62(1), [online], 16–24. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3083.2005.01639.x> [accessed 9 Sep 2021].
- AIRES, Julio Ramos, Thilo KÖHLER, Hiroshi NIKAIDO and Patrick PLÉSIAT. 1999. 'Involvement of an Active Efflux System in the Natural Resistance of *Pseudomonas Aeruginosa* to Aminoglycosides'. *Antimicrobial Agents and Chemotherapy* 43(11), [online], 2624–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC89534/> [accessed 7 Sep 2021].
- ALEKSHUN, M N and S B LEVY. 1997. 'Regulation of Chromosomally Mediated Multiple Antibiotic Resistance: The Mar Regulon.' *Antimicrobial Agents and Chemotherapy* 41(10), [online], 2067–75. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC164072/> [accessed 7 Sep 2021].
- ALEXANDER, D. L. J., A. TROPSHA and David A. WINKLER. 2015. 'Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models'. *Journal of chemical information and modeling* 55(7), [online], 1316–22. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4530125/> [accessed 6 Oct 2021].
- AL-SAEEDI, Mashael and Sahal AL-HAJOJ. 2017. 'Diversity and Evolution of Drug Resistance Mechanisms in *Mycobacterium tuberculosis*'. *Infection and Drug Resistance* 10, [online], 333–42. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648319/> [accessed 25 Jan 2021].
- ANDO, Hiroki, Tohru MIYOSHI-AKIYAMA, Shinya WATANABE and Teruo KIRIKAE. 2014. 'A Silent Mutation in MabA Confers Isoniazid Resistance on *Mycobacterium tuberculosis*'. *Molecular Microbiology* 91(3), [online], 538–47. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/mmi.12476> [accessed 8 Sep 2021].
- ANDREWS, S. 2022. 'Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data'. [online]. Available at: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> [accessed 27 Apr 2022].

- ANDRIES, Koen et al. 2014. 'Acquired Resistance of *Mycobacterium tuberculosis* to Bedaquiline'. *PLoS ONE* 9(7), [online]. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4092087/> [accessed 25 Jan 2021].
- ARAVINDHAN, Vivekanandan et al. 2009. '*Mycobacterium tuberculosis* GroE Promoter Controls the Expression of the Bicistronic GroESL1 Operon and Shows Differential Regulation under Stress Conditions'. *FEMS Microbiology Letters* 292(1), [online], 42–9. Available at: <https://academic.oup.com/femsle/article/292/1/42/490157> [accessed 10 Nov 2020].
- ARNVIG, Kristine B. et al. 2011. 'Sequence-Based Analysis Uncovers an Abundance of Non-Coding RNA in the Total Transcriptome of *Mycobacterium tuberculosis*'. *PLOS Pathog* 7(11), [online], e1002342. Available at: <http://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1002342> [accessed 10 Aug 2016].
- ARNVIG, Kristine and Douglas YOUNG. 2012. 'Non-Coding RNA and Its Potential Role in *Mycobacterium tuberculosis* Pathogenesis'. *RNA Biology* 9(4), [online], 427–36. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3384566/> [accessed 7 Aug 2016].
- BAGCHI, Gargi, Santosh CHAUHAN, Deepak SHARMA and Jaya SivaswamiYR 2005 TYAGI. 2005. 'Transcription and Autoregulation of the Rv3134c-DevR-DevS Operon of *Mycobacterium tuberculosis*'. *Microbiology* 151(12), [online], 4045–53. Available at: <https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.28333-0> [accessed 9 Sep 2021].
- BAH, Aïcha et al. 2020. 'The Lipid Virulence Factors of *Mycobacterium tuberculosis* Exert Multilayered Control over Autophagy-Related Pathways in Infected Human Macrophages'. *Cells* 9(3), E666.
- BAILEY, Timothy L., Nadya WILLIAMS, Chris MISLEH and Wilfred W. LI. 2006. 'MEME: Discovering and Analyzing DNA and Protein Sequence Motifs'. *Nucleic Acids Research* 34(suppl_2), [online], W369–73. Available at: <https://doi.org/10.1093/nar/gkl198> [accessed 31 Jan 2023].
- BAILO, Rebeca, Apoorva BHATT and José A. AÍNSA. 2015. 'Lipid Transport in *Mycobacterium tuberculosis* and Its Implications in Virulence and Drug Development'. *Biochemical Pharmacology* 96(3), [online], 159–67. Available at:

<https://www.sciencedirect.com/science/article/pii/S0006295215002476>
[accessed 14 Oct 2021].

- BANERJEE, A. et al. 1994. 'InhA, a Gene Encoding a Target for Isoniazid and Ethionamide in *Mycobacterium tuberculosis*'. *Science (New York, N.Y.)* 263(5144), 227–30.
- BANERJEE, Asesh, Michele SUGANTINO, James C. SACCHETTINI and William R. JACOBS. 1998. 'The MbaA Gene from the InhA Operon of *Mycobacterium tuberculosis* Encodes a 3-ketoacyl Reductase That Fails to Confer Isoniazid Resistance'. *Microbiology* 144(10), [online], 2697–704. Available at:
<https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-144-10-2697> [accessed 14 Sep 2021].
- BARANOVA, Natalya and Hiroshi NIKAIDO. 2002. 'The BaeSR Two-Component Regulatory System Activates Transcription of the YegMNOB (MdtABCD) Transporter Gene Cluster in *Escherichia coli* and Increases Its Resistance to Novobiocin and Deoxycholate'. *Journal of Bacteriology* 184(15), [online], 4168–76. Available at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC135214/> [accessed 7 Sep 2021].
- BASHYAM, M. D., D. KAUSHAL, S. K. DASGUPTA and A. K. TYAGI. 1996. 'A Study of Mycobacterial Transcriptional Apparatus: Identification of Novel Features in Promoter Elements'. *Journal of Bacteriology* 178(16), 4847–53.
- BELIN, Dominique. 1996. 'The RNase Protection Assay'. In Adrian J. HARWOOD (ed.). *Basic DNA and RNA Protocols*. Totowa, NJ: Humana Press, 131–6. Available at: <https://doi.org/10.1385/0-89603-402-X:131> [accessed 14 Sep 2021].
- BELLEROSE, Michelle M. et al. 2019. 'Common Variants in the Glycerol Kinase Gene Reduce Tuberculosis Drug Efficacy'. *mBio* 10(4), [online], e00663-19. Available at: <https://journals.asm.org/doi/full/10.1128/mBio.00663-19> [accessed 14 Sep 2021].
- BENJAMINI, Yoav and Yosef HOCHBERG. 1995. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1), [online], 289–300. Available at: <http://www.jstor.org/stable/2346101>.
- BENNETT, H. P. J. and D. J. CLARKE. 2005. 'The PbgPE Operon in *Photobacterium luminescens* Is Required for Pathogenicity and Symbiosis'. *Journal of Bacteriology* 187(1), [online], 77–84. Available at:

<https://journals.asm.org/doi/10.1128/jb.187.1.77-84.2005> [accessed 7 Sep 2021].

BERGMAN, Nicholas H., Karla D. PASSALACQUA, Philip C. HANNA and Zhaohui S. QIN. 2007. 'Operon Prediction for Sequenced Bacterial Genomes without Experimental Information'. *Applied and Environmental Microbiology* 73(3), [online], 846–54. Available at: <https://aem.asm.org/content/73/3/846> [accessed 11 Dec 2019].

BHAT, Aadil H., Deepika PATHAK and Alka RAO. 2017. 'The Alr-GroEL1 Operon in *Mycobacterium tuberculosis* : An Interplay of Multiple Regulatory Elements'. *Scientific Reports* 7(1), [online], 43772. Available at: <https://www.nature.com/articles/srep43772> [accessed 10 Nov 2020].

BHATT, Apoorva et al. 2005. 'Conditional Depletion of KasA, a Key Enzyme of Mycolic Acid Biosynthesis, Leads to Mycobacterial Cell Lysis'. *Journal of Bacteriology* 187(22), [online], 7596–606. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1280301/> [accessed 9 Sep 2021].

BOCKHORST, Joseph et al. 2003. 'A Bayesian Network Approach to Operon Prediction'. *Bioinformatics* 19(10), [online], 1227–35. Available at: <https://academic.oup.com/bioinformatics/article/19/10/1227/184417/A-Bayesian-network-approach-to-operon-prediction> [accessed 25 Aug 2017].

BOIROJU, Naveen. 2011. 'A Bootstrap Test for Equality of Mean Absolute Errors'. *ARPN Journal of Engineering and Applied Sciences*.

BOLGER, Anthony M., Marc LOHSE and Bjoern USADEL. 2014. 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data'. *Bioinformatics* 30(15), [online], 2114–20. Available at: <https://doi.org/10.1093/bioinformatics/btu170> [accessed 23 Mar 2022].

BRENDEL, V. and E. N. TRIFONOV. 1984. 'A Computer Algorithm for Testing Potential Prokaryotic Terminators'. *Nucleic Acids Research* 12(10), 4411–27.

BRETL, Daniel J. et al. 2012. 'MprA and DosR Coregulate a *Mycobacterium tuberculosis* Virulence Operon Encoding Rv1813c and Rv1812c'. *Infection and Immunity* 80(9), [online], 3018–33. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3418728/> [accessed 29 Nov 2016].

- BRIFFOTAUX, Julien, Wei HUANG, Xinwei WANG and Brigitte GICQUEL. 2017. 'MmpS5/MmpL5 as an Efflux Pump in Mycobacterium Species'. *Tuberculosis (Edinburgh, Scotland)* 107, 13–9.
- BROGNA, S and M ASHBURNER. 1997. 'The Adh-Related Gene of *Drosophila Melanogaster* Is Expressed as a Functional Dicistronic Messenger RNA: Multigenic Transcription in Higher Organisms.' *The EMBO Journal* 16(8), [online], 2023–31. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1169805/> [accessed 26 Jan 2021].
- BROOUN, Alexei, John J. TOMASHEK and Kim LEWIS. 1999. 'Purification and Ligand Binding of EmrR, a Regulator of a Multidrug Transporter'. *Journal of Bacteriology* 181(16), [online], 5131–3. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC94011/> [accessed 7 Sep 2021].
- BUNDALOVIC-TORMA, Cedoljub et al. 2020. 'A Systematic Pipeline for Classifying Bacterial Operons Reveals the Evolutionary Landscape of Biofilm Machineries'. *PLOS Computational Biology* 16(4), [online], e1007721. Available at: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007721> [accessed 17 Sep 2021].
- CAO, Huansheng, Qin MA, Xin CHEN and Ying XU. 2019. 'DOOR: A Prokaryotic Operon Database for Genome Analyses and Functional Inference'. *Briefings in Bioinformatics* 20(4), 1568–77.
- CASALI, Nicola et al. 2016. 'Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study'. *PLOS Med* 13(10), [online], e1002137. Available at: <http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002137> [accessed 4 Nov 2016].
- CASART, Yveth et al. 2008. 'Par Genes in *Mycobacterium Bovis* and *Mycobacterium Smegmatis* Are Arranged in an Operon Transcribed from "SigGC" Promoters'. *BMC Microbiology* 8, [online], 51. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2346475/> [accessed 9 Sep 2021].
- CASTRO, Rhastin A D et al. 2020. 'The Genetic Background Modulates the Evolution of Fluoroquinolone-Resistance in *Mycobacterium tuberculosis*'. *Molecular Biology and Evolution* 37(1), [online], 195–207. Available at: <https://doi.org/10.1093/molbev/msz214> [accessed 14 Sep 2021].

- CDCTB. 2016. 'MDRTB Factsheet'. *Centers for Disease Control and Prevention* [online]. Available at: <https://www.cdc.gov/tb/publications/factsheets/drtb/mdrtb.htm> [accessed 30 Jan 2023].
- CHAN, Raphael C. Y. et al. 2007. 'Genetic and Phenotypic Characterization of Drug-Resistant *Mycobacterium tuberculosis* Isolates in Hong Kong'. *Journal of Antimicrobial Chemotherapy* 59(5), [online], 866–73. Available at: <http://jac.oxfordjournals.org/content/59/5/866> [accessed 15 Feb 2016].
- CHE, Dongsheng et al. 2006. 'Detecting Uber-Operons in Prokaryotic Genomes'. *Nucleic Acids Research* 34(8), [online], 2418–27. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1458513/> [accessed 11 Dec 2019].
- CHEN, Xin, Zhengchang SU, Ying XU and Tao JIANG. 2004. 'Computational Prediction of Operons in *Synechococcus* Sp. WH8102'. *Genome Informatics. International Conference on Genome Informatics* 15(2), 211–22.
- CHEON, Hyejin. 2017. 'Comparison of CT Findings of between MDR-TB and XDR-TB: A Propensity Score Matching Study'. *Imaging in Medicine* 9(5), [online], 125–30. Available at: <https://www.openaccessjournals.com/abstract/comparison-of-ct-findings-of-between-mdrtb-and-xdrtb-a-propensity-score-matching-study-12197.html> [accessed 30 Jan 2023].
- CHUANG, Li-Yeh, Hsueh-Wei CHANG, Jui-Hung TSAI and Cheng-Hong YANG. 2012. 'Features for Computational Operon Prediction in Prokaryotes'. *Briefings in Functional Genomics* 11(4), [online], 291–9. Available at: <https://academic.oup.com/bfg/article/11/4/291/200556> [accessed 30 Jan 2020].
- COLANGELI, Roberto et al. 2005. 'The *Mycobacterium tuberculosis* IniA Gene Is Essential for Activity of an Efflux Pump That Confers Drug Tolerance to Both Isoniazid and Ethambutol'. *Molecular Microbiology* 55(6), [online], 1829–40. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2958.2005.04510.x> [accessed 7 Sep 2021].
- COLANGELI, Roberto et al. 2014. 'Whole Genome Sequencing of *Mycobacterium tuberculosis* Reveals Slow Growth and Low Mutation Rates during Latent Infections in Humans'. Edited by Deepak Kaushal. *PLoS ONE* 9(3), [online], e91024. Available at: <https://dx.plos.org/10.1371/journal.pone.0091024> [accessed 7 Jul 2022].

- COLE, S. T. et al. 1998. 'Deciphering the Biology of *Mycobacterium tuberculosis* from the Complete Genome Sequence'. *Nature* 393(6685), [online], 537–44. Available at: <http://www.nature.com/nature/journal/v393/n6685/full/393537a0.html> [accessed 16 Feb 2016].
- COLL, Francesc et al. 2018. 'Genome-Wide Analysis of Multi- and Extensively Drug-Resistant *Mycobacterium tuberculosis*'. *Nature Genetics* 50(2), [online], 307–16. Available at: <https://www.nature.com/articles/s41588-017-0029-0> [accessed 17 Apr 2020].
- COMAS, Iñaki et al. 2012. 'Whole-Genome Sequencing of Rifampicin-Resistant *Mycobacterium tuberculosis* Strains Identifies Compensatory Mutations in RNA Polymerase Genes'. *Nature Genetics* 44(1), [online], 106–10. Available at: <https://www.nature.com/articles/ng.1038> [accessed 16 Sep 2021].
- COMAS, Iñaki et al. 2013. 'Out-of-Africa Migration and Neolithic Co-Expansion of *Mycobacterium tuberculosis* with Modern Humans'. *Nature genetics* 45(10), [online], 1176–82. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3800747/> [accessed 25 Jan 2021].
- COMAS, Iñaki et al. 2015. 'Population Genomics of *Mycobacterium tuberculosis* in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa'. *Current Biology* 25(24), [online], 3260–6. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4691238/> [accessed 14 Sep 2020].
- CONESA, Ana et al. 2016. 'A Survey of Best Practices for RNA-Seq Data Analysis'. *Genome Biology* 17(1), [online], 13. Available at: <https://doi.org/10.1186/s13059-016-0881-8> [accessed 23 Feb 2023].
- CORTES, Teresa et al. 2013. 'Genome-Wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in *Mycobacterium tuberculosis*'. *Cell Reports* 5(4), 1121–31.
- COSCOLLA, Mireia et al. 2015. 'M. Tuberculosis T Cell Epitope Analysis Reveals Paucity of Antigenic Variation and Identifies Rare Variable TB Antigens'. *Cell Host & Microbe* 18(5), [online], 538–48. Available at: <http://www.sciencedirect.com/science/article/pii/S1931312815004187> [accessed 16 Feb 2016].
- COSCOLLA, Mireia and Sebastien GAGNEUX. 2014. 'Consequences of Genomic Diversity in *Mycobacterium tuberculosis*'. *Seminars in*

Immunology 26(6), [online], 431–44. Available at:
<http://www.sciencedirect.com/science/article/pii/S1044532314000967>
[accessed 9 Nov 2016].

DAM, Phuongan et al. 2007. ‘Operon Prediction Using Both Genome-Specific and General Genomic Information’. *Nucleic Acids Research* 35(1), [online], 288–98. Available at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1802555/> [accessed 4 Sep 2021].

DANDEKAR, T., B. SNEL, M. HUYNEN and P. BORK. 1998. ‘Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact’. *Trends in Biochemical Sciences* 23(9), 324–8.

DERZELLE, Sylviane et al. 2004. ‘The PhoP-PhoQ Two-Component Regulatory System of *Photobacterium luminescens* Is Essential for Virulence in Insects’. *Journal of Bacteriology* 186(5), [online], 1270–9. Available at:
<https://journals.asm.org/doi/10.1128/jb.186.5.1270-1279.2004> [accessed 7 Sep 2021].

DEVRIES, Zachary et al. 2021. ‘Using a National Surgical Database to Predict Complications Following Posterior Lumbar Surgery and Comparing the Area under the Curve and F1-Score for the Assessment of Prognostic Capability’. *The Spine Journal: Official Journal of the North American Spine Society* 21(7), 1135–42.

DOIG, C, A L SEAGAR, B WATT and K J FORBES. 2002. ‘The Efficacy of the Heat Killing of *Mycobacterium tuberculosis*’. *Journal of Clinical Pathology* 55(10), [online], 778–9. Available at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1769777/> [accessed 25 Jan 2021].

ERMOLAEVA, M. D., O. WHITE and S. L. SALZBERG. 2001. ‘Prediction of Operons in Microbial Genomes’. *Nucleic Acids Research* 29(5), 1216–21.

EVANS, Kelly, Lateef ADEWOYE and Keith POOLE. 2001. ‘MexR Repressor of the MexAB-OprM Multidrug Efflux Operon of *Pseudomonas aeruginosa*: Identification of MexR Binding Sites in the MexA-MexR Intergenic Region’. *Journal of Bacteriology* 183(3), [online], 807–12. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC94945/> [accessed 7 Sep 2021].

FORD, Christopher B. et al. 2013. ‘*Mycobacterium tuberculosis* Mutation Rate Estimates from Different Lineages Predict Substantial Differences in the Emergence of Drug-Resistant Tuberculosis’. *Nature Genetics* 45(7),

[online], 784–90. Available at: <https://www.nature.com/articles/ng.2656> [accessed 25 Jan 2021].

FORTINO, Vittorio et al. 2014. ‘Transcriptome Dynamics-Based Operon Prediction in Prokaryotes’. *BMC Bioinformatics* 15(1), [online], 145. Available at: <https://doi.org/10.1186/1471-2105-15-145> [accessed 11 Dec 2019].

FU, Li M. 2006. ‘Exploring Drug Action on *Mycobacterium tuberculosis* Using Affymetrix Oligonucleotide Genechips’. *Tuberculosis (Edinburgh, Scotland)* 86(2), [online], 134–43. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1557687/> [accessed 7 Jun 2022].

FU, Li M. and Shu C. TAI. 2009. ‘The Differential Gene Expression Pattern of *Mycobacterium tuberculosis* in Response to Capreomycin and PA-824 versus First-Line TB Drugs Reveals Stress- and PE/PPE-Related Drug Targets’. *International Journal of Microbiology* 2009, [online], 879621. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2775200/> [accessed 10 Oct 2021].

GAGNEUX, Sebastien et al. 2006. ‘The Competitive Cost of Antibiotic Resistance in *Mycobacterium tuberculosis*’. *Science* 312(5782), [online], 1944–6. Available at: <https://science.sciencemag.org/content/312/5782/1944> [accessed 25 Jan 2021].

GARIMA, Kushal et al. 2015. ‘Differential Expression of Efflux Pump Genes of *Mycobacterium tuberculosis* in Response to Varied Subinhibitory Concentrations of Antituberculosis Agents’. *Tuberculosis (Edinburgh, Scotland)* 95(2), 155–61.

GIOFFRÉ, Andrea et al. 2005. ‘Mutation in Mce Operons Attenuates *Mycobacterium tuberculosis* Virulence’. *Microbes and Infection* 7(3), [online], 325–34. Available at: <https://www.sciencedirect.com/science/article/pii/S1286457905000092> [accessed 14 Sep 2021].

GOUDE, Renan, Anita G. AMIN, Delphi CHATTERJEE and Tanya PARISH. 2008. ‘The Critical Role of EmbC in *Mycobacterium tuberculosis*’. *Journal of Bacteriology* 190(12), [online], 4335–41. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2446762/> [accessed 1 Oct 2021].

GRKOVIC, Steve et al. 2001. ‘The Staphylococcal QacR Multidrug Regulator Binds a Correctly Spaced Operator as a Pair of Dimers’. *Journal of*

Bacteriology 183(24), [online], 7102–9. Available at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC95558/> [accessed 7 Sep 2021].

GÜELL, Marc et al. 2009. ‘Transcriptome Complexity in a Genome-Reduced Bacterium’. *Science (New York, N.Y.)* 326(5957), 1268–71.

HALLER, Rachel, Meghann KENNEDY, Nick ARNOLD and Robert RUTHERFORD. 2010. ‘The Transcriptome of *Mycobacterium tuberculosis*’. *Applied Microbiology and Biotechnology* 86(1), 1–9.

HOAGLAND, Daniel T., Jiuyu LIU, Robin B. LEE and Richard E. LEE. 2016. ‘New Agents for the Treatment of Drug-Resistant *Mycobacterium tuberculosis*’. *Advanced Drug Delivery Reviews* 102, [online], 55–72. Available at:
<http://www.sciencedirect.com/science/article/pii/S0169409X16301363> [accessed 17 Apr 2020].

HUNT, Debbie M. et al. 2012. ‘Long-Range Transcriptional Control of an Operon Necessary for Virulence-Critical ESX-1 Secretion in *Mycobacterium tuberculosis*’. *Journal of Bacteriology* 194(9), 2307–20.

HUNTER, John D. 2007. ‘Matplotlib: A 2D Graphics Environment’. *Computing in Science and Engineering* 9(3), [online], 90–5. Available at:
<https://doi.org/10.1109/MCSE.2007.55> [accessed 23 Mar 2022].

HUSTMYER, Christine M. et al. 2018. ‘Promoter Boundaries for the LuxCDABE and BetIBA-ProXWV Operons in *Vibrio Harveyi* Defined by the Method Rapid Arbitrary PCR Insertion Libraries (RAIL)’. *Journal of Bacteriology* 200(11), [online], e00724-17. Available at:
<https://journals.asm.org/doi/full/10.1128/JB.00724-17> [accessed 13 Sep 2021].

JACOB, E., R. SASIKUMAR and K. N. R. NAIR. 2005. ‘A Fuzzy Guided Genetic Algorithm for Operon Prediction’. *Bioinformatics (Oxford, England)* 21(8), 1403–7.

JACOB, F., D. PERRIN, C. SANCHEZ and J. MONOD. 1960. ‘[Operon: a group of genes with the expression coordinated by an operator]’. *Comptes Rendus Hebdomadaires Des Seances De l’Academie Des Sciences* 250, 1727–9.

JACOB, François and Jacques MONOD. 1961. ‘Genetic Regulatory Mechanisms in the Synthesis of Proteins’. *Journal of Molecular Biology* 3(3), [online], 318–56. Available at:

<https://www.sciencedirect.com/science/article/pii/S0022283661800727>
[accessed 2 Sep 2021].

JALILI, Vahid et al. 2020. 'The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2020 Update'. *Nucleic Acids Research* 48(W1), W395–402.

JANGA, Sarath Chandra, Julio COLLADO-VIDES and Gabriel MORENO-HAGELSIEB. 2005. 'Nebulon: A System for the Inference of Functional Relationships of Gene Products from the Rearrangement of Predicted Operons'. *Nucleic Acids Research* 33(8), [online], 2521–30. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1088069/> [accessed 2 Sep 2021].

JENSEN, Lars J. et al. 2009. 'STRING 8--a Global View on Proteins and Their Functional Interactions in 630 Organisms'. *Nucleic Acids Research* 37(Database issue), D412-416.

KAPOPOULOU, Adamandia, Jocelyne M. LEW and Stewart T. COLE. 2011. 'The MycoBrowser Portal: A Comprehensive and Manually Annotated Resource for Mycobacterial Genomes'. *Tuberculosis (Edinburgh, Scotland)* 91(1), 8–13.

KHAW, Khai Wah, Xinying CHEW, Wai Chung YEONG and Sok Li LIM. 2019. 'Optimal Design of the Synthetic Control Chart for Monitoring the Multivariate Coefficient of Variation'. *Chemometrics and Intelligent Laboratory Systems* 186, [online], 33–40. Available at: <https://www.sciencedirect.com/science/article/pii/S0169743918304726> [accessed 4 Feb 2023].

KIEBOOM, Jasper and Jan A. M. YR 2001 DE BONT. 2001. 'Identification and Molecular Characterization of an Efflux System Involved in *Pseudomonas Putida* S12 Multidrug Resistance'. *Microbiology* 147(1), [online], 43–51. Available at: <https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-147-1-43> [accessed 7 Sep 2021].

KLUYVER, Thomas et al. 2016. 'Jupyter Notebooks – a Publishing Format for Reproducible Computational Workflows'. In Fernando LOIZIDES and Birgit SCMIDT (eds.). 20th International Conference on Electronic Publishing (01/01/16), 2016, 87–90. Available at: <https://eprints.soton.ac.uk/403913/> [accessed 27 Apr 2022].

KÖHLER, Thilo, Simone F. EPP, Lasta Kocjancic CURTY and Jean-Claude PECHÈRE. 1999. 'Characterization of MexT, the Regulator of the MexE-MexF-OprN Multidrug Efflux System of *Pseudomonas Aeruginosa*'.

Journal of Bacteriology 181(20), [online], 6300–5. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC103763/> [accessed 7 Sep 2021].

KRISHNAKUMAR, Raga and Anne M. RUFFING. 2022. ‘OperonSEQer: A Set of Machine-Learning Algorithms with Threshold Voting for Detection of Operon Pairs Using Short-Read RNA-Sequencing Data’. *PLoS Computational Biology* 18(1), [online], e1009731. Available at: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009731> [accessed 27 Feb 2023].

LAINING, Emma, Khushwant SIDHU and Simon J. HUBBARD. 2008. ‘Predicted Transcription Factor Binding Sites as Predictors of Operons in Escherichia Coli and Streptomyces Coelicolor’. *BMC genomics* 9, 79.

LAM, T. H. J. et al. 2008. ‘Differential FadE28 Expression Associated with Phenotypic Virulence of *Mycobacterium tuberculosis*’. *Microbial Pathogenesis* 45(1), 12–7.

LAMBERT, R.J.W. and J. PEARSON. 2000. ‘Susceptibility Testing: Accurate and Reproducible Minimum Inhibitory Concentration (MIC) and Non-inhibitory Concentration (NIC) Values’. *Journal of Applied Microbiology* 88(5), [online], 784–90. Available at: <https://doi.org/10.1046/j.1365-2672.2000.01017.x> [accessed 3 Feb 2023].

LAMICHHANE, Shree. 2018. ‘*Mycobacterium tuberculosis*: Gene and Genome Analysis’.

LAW, Charity W, Yunshun CHEN, Wei SHI and Gordon K SMYTH. 2014. ‘Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts’. *Genome Biology* 15(2), [online], R29. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053721/> [accessed 17 Jun 2022].

LEE, Eun-Jin, Jeongjoon CHOI and Eduardo A. GROISMAN. 2014. ‘Control of a Salmonella Virulence Operon by Proline-Charged TRNAPro’. *Proceedings of the National Academy of Sciences* 111(8), [online], 3140–5. Available at: <https://www.pnas.org/content/111/8/3140> [accessed 2 Sep 2021].

LEE, M. H., L. PASCOPELLA, W. R. JACOBS and G. F. HATFULL. 1991. ‘Site-Specific Integration of Mycobacteriophage L5: Integration-Proficient Vectors for Mycobacterium Smegmatis, *Mycobacterium tuberculosis*, and Bacille Calmette-Guérin.’ *Proceedings of the National Academy of Sciences* 88(8), [online], 3111–5. Available at: <https://www.pnas.org/content/88/8/3111> [accessed 16 Sep 2021].

- LEIMKÜHLER, Silke. 2014. 'The Biosynthesis of the Molybdenum Cofactor in Escherichia Coli and Its Connection to FeS Cluster Assembly and the Thiolation of TRNA'. *Advances in Biology* 2014, [online], e808569. Available at: <https://www.hindawi.com/journals/ab/2014/808569/> [accessed 1 Oct 2021].
- LI, Heng et al. 2009. 'The Sequence Alignment/Map Format and SAMtools'. *Bioinformatics* 25(16), [online], 2078–9. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/> [accessed 23 Mar 2022].
- LI, Heng and Richard DURBIN. 2009. 'Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform'. *Bioinformatics* 25(14), [online], 1754–60. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705234/> [accessed 23 Mar 2022].
- LI, Jin. 2017. 'Assessing the Accuracy of Predictive Models for Numerical Data: Not r nor R2, Why Not? Then What?' *PLOS ONE* 12(8), [online], e0183250. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183250> [accessed 6 Oct 2021].
- LI, Xian-Zhi, Li ZHANG and Keith POOLE. 2002. 'SmeC, an Outer Membrane Multidrug Efflux Protein of *Stenotrophomonas Maltophilia*'. *Antimicrobial Agents and Chemotherapy* 46(2), [online], 333–43. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC127032/> [accessed 7 Sep 2021].
- LIU, Ruijie et al. 2015. 'Why Weight? Modelling Sample and Observational Level Variability Improves Power in RNA-Seq Analyses'. *Nucleic Acids Research* 43(15), [online], e97. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4551905/> [accessed 17 Jun 2022].
- LOMOVSKAYA, O, F KAWAI and A MATIN. 1996. 'Differential Regulation of the Mcb and Emr Operons of Escherichia Coli: Role of Mcb in Multidrug Resistance.' *Antimicrobial Agents and Chemotherapy* 40(4), [online], 1050–2. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC163261/> [accessed 7 Sep 2021].
- LOMOVSKAYA, O., K. LEWIS and A. MATIN. 1995. 'EmrR Is a Negative Regulator of the Escherichia Coli Multidrug Resistance Pump EmrAB'. *Journal of Bacteriology* 177(9), 2328–34.

- LUCAS, C E, J T BALTHAZAR, K E HAGMAN and W M SHAFER. 2014. 'The MtrR Repressor Binds the DNA Sequence between the MtrR and MtrC Genes of *Neisseria Gonorrhoeae*.' *Journal of Bacteriology* 179(13), [online], 4123–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC179230/> [accessed 7 Sep 2021].
- LYNCH, D. et al. 2001. 'Genetic Organization of the Region Encoding Regulation, Biosynthesis, and Transport of Rhizobactin 1021, a Siderophore Produced by *Sinorhizobium Meliloti*'. *Journal of Bacteriology* 183(8), 2576–85.
- MAGNET, Sophie, Patrice COURVALIN and Thierry LAMBERT. 2001. 'Resistance-Nodulation-Cell Division-Type Efflux Pump Involved in Aminoglycoside Resistance in *Acinetobacter Baumannii* Strain BM4454'. *Antimicrobial Agents and Chemotherapy* 45(12), [online], 3375–80. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC90840/> [accessed 7 Sep 2021].
- MAHONY, James B. and Max A. CHERNESKY. 1995. '10 - Multiplex Polymerase Chain Reaction'. In Danny L. WIEDBRAUK and Daniel H. FARKAS (eds.). *Molecular Methods for Virus Detection*. San Diego: Academic Press, 219–36. Available at: <https://www.sciencedirect.com/science/article/pii/B978012748920950011X> [accessed 8 Sep 2021].
- MANGANELLI, Riccardo et al. 2004a. 'σ Factors and Global Gene Regulation in *Mycobacterium tuberculosis*'. *Journal of Bacteriology* 186(4), [online], 895–902. Available at: <http://jb.asm.org/content/186/4/895> [accessed 29 Nov 2016].
- MANGANELLI, Riccardo et al. 2004b. 'σ Factors and Global Gene Regulation in *Mycobacterium tuberculosis*'. *Journal of Bacteriology* 186(4), [online], 895–902. Available at: <http://jb.asm.org/content/186/4/895> [accessed 29 Nov 2016].
- MAO, Fenglou et al. 2009. 'DOOR: A Database for Prokaryotic Operons'. *Nucleic Acids Research* 37(suppl_1), [online], D459–63. Available at: https://academic.oup.com/nar/article/37/suppl_1/D459/1008910 [accessed 31 Jan 2020].
- MARIAM, Deneke H., Yohannes MENGISTU, Sven E. HOFFNER and Dan I. ANDERSSON. 2004. 'Effect of RpoB Mutations Conferring Rifampin Resistance on Fitness of *Mycobacterium tuberculosis*'. *Antimicrobial Agents and Chemotherapy* 48(4), [online], 1289–94. Available at:

<https://journals.asm.org/doi/10.1128/AAC.48.4.1289-1294.2004> [accessed 16 Sep 2021].

MCCARTHY, Davis J. and Gordon K. SMYTH. 2009. 'Testing Significance Relative to a Fold-Change Threshold Is a TREAT'. *Bioinformatics* 25(6), [online], 765–71. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2654802/> [accessed 17 Jun 2022].

MCCLURE, Ryan et al. 2013. 'Computational Analysis of Bacterial RNA-Seq Data'. *Nucleic Acids Research* 41(14), [online], e140. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3737546/> [accessed 23 Oct 2019].

MENDEL, Ralf R. 2013. 'The Molybdenum Cofactor'. *The Journal of Biological Chemistry* 288(19), [online], 13165–72. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3650355/> [accessed 1 Oct 2021].

MILANO, Anna et al. 2009. 'Azole Resistance in *Mycobacterium tuberculosis* Is Mediated by the MmpS5–MmpL5 Efflux System'. *Tuberculosis* 89(1), [online], 84–90. Available at: <https://www.sciencedirect.com/science/article/pii/S1472979208000942> [accessed 27 Jun 2022].

MINER, Maurine D. et al. 2009. 'Role of Cholesterol in *Mycobacterium tuberculosis* Infection'. *Indian Journal of Experimental Biology* 47(6), 407–11.

MIRYALA, Sravan Kumar, Anand ANBARASU and Sudha RAMAIAH. 2019. 'Impact of Bedaquiline and Capreomycin on the Gene Expression Patterns of Multidrug-Resistant *Mycobacterium tuberculosis* H37Rv Strain and Understanding the Molecular Mechanism of Antibiotic Resistance'. *Journal of Cellular Biochemistry* 120(9), [online], 14499–509. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcb.28711> [accessed 14 Sep 2020].

MOORE, Richard A. et al. 1999. 'Efflux-Mediated Aminoglycoside and Macrolide Resistance in *Burkholderia Pseudomallei*'. *Antimicrobial Agents and Chemotherapy* 43(3), [online], 465–70. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC89145/> [accessed 7 Sep 2021].

MOUAMMINE, Annabelle et al. 2017. 'An Antimicrobial Peptide-Resistant Minor Subpopulation of *Photobacterium luminescens* Is Responsible for

Virulence'. *Scientific Reports* 7(1), [online], 43670. Available at: <https://www.nature.com/articles/srep43670> [accessed 7 Sep 2021].

MUSTAFA, Abu S. 2011. 'Comparative Evaluation of MPT83 (Rv2873) for T Helper-1 Cell Reactivity and Identification of HLA-Promiscuous Peptides in Mycobacterium Bovis BCG-Vaccinated Healthy Subjects ▽'. *Clinical and Vaccine Immunology : CVI* 18(10), [online], 1752–9. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3187038/> [accessed 7 Jun 2022].

MUTZ, Kai-Oliver et al. 2013. 'Transcriptome Analysis Using Next-Generation Sequencing'. *Current Opinion in Biotechnology* 24(1), [online], 22–30. Available at: <https://www.sciencedirect.com/science/article/pii/S0958166912001310> [accessed 13 Sep 2021].

NAGAKUBO, Satoshi, Kunihiko NISHINO, Takahiro HIRATA and Akihito YAMAGUCHI. 2002. 'The Putative Response Regulator BaeR Stimulates Multidrug Resistance of Escherichia Coli via a Novel Multidrug Exporter System, MdtABC'. *Journal of Bacteriology* 184(15), [online], 4161–7. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC135206/> [accessed 7 Sep 2021].

NAIDOO, Charissa C. and Manormoney PILLAY. 2017. 'Fitness-Compensatory Mutations Facilitate the Spread of Drug-Resistant F15/LAM4/KZN and F28 *Mycobacterium tuberculosis* Strains in KwaZulu-Natal, South Africa'. *Journal of Genetics* 96(4), [online], 599–612. Available at: <https://doi.org/10.1007/s12041-017-0805-8> [accessed 16 Sep 2021].

NAMOUCHE, Amine et al. 2016. 'The *Mycobacterium tuberculosis* Transcriptional Landscape under Genotoxic Stress'. *BMC Genomics* 17, [online], 791. Available at: <http://dx.doi.org/10.1186/s12864-016-3132-1> [accessed 3 Feb 2017].

NARANG, Anshika et al. 2019. 'Potential Impact of Efflux Pump Genes in Mediating Rifampicin Resistance in Clinical Isolates of *Mycobacterium tuberculosis* from India'. *PloS One* 14(9), e0223163.

NAVILLE, Magali and Daniel GAUTHERET. 2009. 'Transcription Attenuation in Bacteria: Theme and Variations'. *Briefings in Functional Genomics* 8(6), [online], 482–92. Available at: <https://doi.org/10.1093/bfgp/elp025> [accessed 30 Jun 2022].

NEBENZAHL-GUIMARAES, Hanna et al. 2016. 'Genomic Characterization of *Mycobacterium tuberculosis* Lineage 7 and a Proposed Name: "Aethiops Vetus"'. *Microbial Genomics* 2(6), [online]. Available at:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320646/> [accessed 25 Jan 2021].

OKUDA, Shujiro et al. 2007. 'Characterization of Relationships between Transcriptional Units and Operon Structures in *Bacillus Subtilis* and *Escherichia Coli*'. *BMC Genomics* 8, [online], 48. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1808063/> [accessed 3 Sep 2021].

ORGEUR, Mickael and Roland BROSCHE. 2018. 'Evolution of Virulence in the *Mycobacterium tuberculosis* Complex'. *Current Opinion in Microbiology* 41, [online], 68–75. Available at: <http://www.sciencedirect.com/science/article/pii/S1369527417300796> [accessed 25 Jan 2021].

OSBOURN, Anne E. and Ben FIELD. 2009. 'Operons'. *Cellular and Molecular Life Sciences* 66(23), [online], 3755–75. Available at: <https://doi.org/10.1007/s00018-009-0114-3> [accessed 19 Oct 2019].

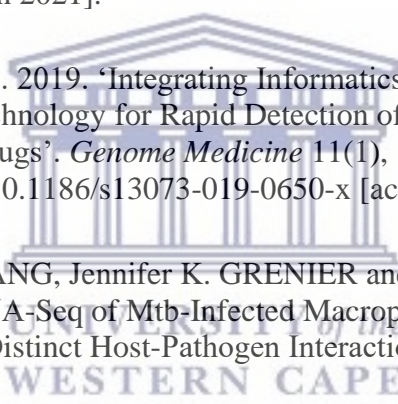
OVERBEEK, Ross et al. 1999. 'The Use of Gene Clusters to Infer Functional Coupling'. *Proceedings of the National Academy of Sciences* 96(6), [online], 2896–901. Available at: <https://www.pnas.org/content/96/6/2896> [accessed 2 Sep 2021].

OZOLINE, O. N., A. A. DEEV and M. V. ARKHIPOVA. 1997. 'Non-Canonical Sequence Elements in the Promoter Structure. Cluster Analysis of Promoters Recognized by *Escherichia Coli* RNA Polymerase'. *Nucleic Acids Research* 25(23), 4703–9.

PAN, W. and B. G. SPRATT. 1994. 'Regulation of the Permeability of the Gonococcal Cell Envelope by the Mtr System'. *Molecular Microbiology* 11(4), 769–75.

PANDEY, Amit K. and Christopher M. SASSETTI. 2008. 'Mycobacterial Persistence Requires the Utilization of Host Cholesterol'. *Proceedings of the National Academy of Sciences of the United States of America* 105(11), [online], 4376–80. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2393810/> [accessed 7 Jul 2022].

PARISH, Tanya et al. 2003. 'Deletion of Two-Component Regulatory Systems Increases the Virulence of *Mycobacterium tuberculosis*'. *Infection and Immunity* 71(3), [online], 1134–40. Available at: <https://iai.asm.org/content/71/3/1134> [accessed 25 Jan 2021].

- PARKER, Nina et al. 2016. 'Gene Regulation: Operon Theory' [online]. Available at: <https://opentextbc.ca/microbiologyopenstax/chapter/gene-regulation-operon-theory/> [accessed 7 Oct 2021].
- PASCA, Maria Rosalia et al. 2004. 'Rv2686c-Rv2687c-Rv2688c, an ABC Fluoroquinolone Efflux Pump in *Mycobacterium tuberculosis*'. *Antimicrobial Agents and Chemotherapy* 48(8), [online], 3175–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC478549/> [accessed 7 Sep 2021].
- PELLY, Shaaretha et al. 2016. 'REMap: Operon Map of *M. Tuberculosis*'. *Tuberculosis (Edinburgh, Scotland)* 99, [online], 70–80. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4967370/> [accessed 19 Oct 2019].
- PETERSON, Eliza J. R. et al. 2020. 'Intricate Genetic Programs Controlling Dormancy in *Mycobacterium tuberculosis*'. *Cell Reports* 31(4), [online], 107577. Available at: <https://www.sciencedirect.com/science/article/pii/S221112472030526X> [accessed 17 Jun 2021].
- PHELAN, Jody E. et al. 2019. 'Integrating Informatics Tools and Portable Sequencing Technology for Rapid Detection of Resistance to Anti-Tuberculous Drugs'. *Genome Medicine* 11(1), [online], 41. Available at: <https://doi.org/10.1186/s13073-019-0650-x> [accessed 27 Apr 2022].
- PISU, Davide, Lu HUANG, Jennifer K. GRENIER and David G. RUSSELL. 2020. 'Dual RNA-Seq of Mtb-Infected Macrophages In Vivo Reveals Ontologically Distinct Host-Pathogen Interactions'. *Cell Reports* 30(2), 335-350.e4. 
- PLINKE, Claudia et al. 2010. 'EmbCAB Sequence Variation among Ethambutol-Resistant *Mycobacterium tuberculosis* Isolates without EmbB306 Mutation'. *Journal of Antimicrobial Chemotherapy* 65(7), [online], 1359–67. Available at: <https://doi.org/10.1093/jac/dkq120> [accessed 8 Sep 2021].
- POOLE, K et al. 1996. 'Expression of the Multidrug Resistance Operon MexA-MexB-OprM in *Pseudomonas Aeruginosa*: MexR Encodes a Regulator of Operon Expression.' *Antimicrobial Agents and Chemotherapy* 40(9), [online], 2021–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC163466/> [accessed 7 Sep 2021].
- PREZIOSO, Stephanie M. et al. 2018. 'Shikimate Induced Transcriptional Activation of Protocatechuate Biosynthesis Genes by QuiR, a LysR-Type

Transcriptional Regulator, in *Listeria Monocytogenes*'. *Journal of Molecular Biology* 430(9), [online], 1265–83. Available at: <https://www.sciencedirect.com/science/article/pii/S0022283618301256> [accessed 13 Sep 2021].

PRICE, Morgan N., Adam P. ARKIN and Eric J. ALM. 2006. 'The Life-Cycle of Operons'. *PLOS Genetics* 2(6), [online], e96. Available at: <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020096> [accessed 19 Oct 2019].

PRICE, Morgan N., Katherine H. HUANG, Eric J. ALM and Adam P. ARKIN. 2005. 'A Novel Method for Accurate Operon Predictions in All Sequenced Prokaryotes'. *Nucleic Acids Research* 33(3), 880–92.

PUNINA, N. V., N. M. MAKRIDAKIS, M. A. REMNEV and A. F. TOPUNOV. 2015. 'Whole-Genome Sequencing Targets Drug-Resistant Bacterial Infections'. *Human Genomics* 9, [online], 19. Available at: <http://dx.doi.org/10.1186/s40246-015-0037-z> [accessed 11 Feb 2016].

R CORE TEAM. 2021. 'R: A Language and Environment for Statistical Computing'. Available at: <https://www.R-project.org>.

RAO, Mohan S. et al. 2019. 'Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies'. *Frontiers in Genetics* 9, [online]. Available at: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00636/full> [accessed 19 Jun 2021].

RIFAT, Dalin et al. 2017. 'In Vitro and in Vivo Fitness Costs Associated with *Mycobacterium tuberculosis* RpoB Mutation H526D'. *Future Microbiology* 12, 753–65.

RITCHIE, Matthew E et al. 2006. 'Empirical Array Quality Weights in the Analysis of Microarray Data'. *BMC Bioinformatics* 7, [online], 261. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1564422/> [accessed 17 Jun 2022].

ROBACK, P. et al. 2007. 'A Predicted Operon Map for *Mycobacterium tuberculosis*'. *Nucleic Acids Research* 35(15), 5085–95.

ROBINSON, Mark D. and Alicia OSHLACK. 2010. 'A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data'. *Genome Biology* 11(3), [online], R25. Available at: <https://doi.org/10.1186/gb-2010-11-3-r25> [accessed 25 Nov 2021].

- ROMERO, P. R. and P. D. KARP. 2004. 'Using Functional and Organizational Information to Improve Genome-Wide Computational Prediction of Transcription Units on Pathway-Genome Databases'. *Bioinformatics (Oxford, England)* 20(5), 709–17.
- ROUQUETTE, C., J. B. HARMON and W. M. SHAFER. 1999. 'Induction of the MtrCDE-Encoded Efflux Pump System of Neisseria Gonorrhoeae Requires MtrA, an AraC-like Protein'. *Molecular Microbiology* 33(3), 651–8.
- SÁENZ-LAHOYA, S. et al. 2019. 'Noncontiguous Operon Is a Genetic Organization for Coordinating Bacterial Gene Expression'. *Proceedings of the National Academy of Sciences* 116(5), [online], 1733–8. Available at: <https://www.pnas.org/content/116/5/1733> [accessed 6 Sep 2021].
- SALGADO, Heladia, Gabriel MORENO-HAGELSIEB, Temple F. SMITH and Julio COLLADO-VIDES. 2000. 'Operons in Escherichia Coli: Genomic Analyses and Predictions'. *Proceedings of the National Academy of Sciences* 97(12), [online], 6652–7. Available at: <https://www.pnas.org/content/97/12/6652> [accessed 29 Jan 2020].
- SALINA, Elena G. et al. 2019. 'Resuscitation of Dormant "Non-Culturable" *Mycobacterium tuberculosis* Is Characterized by Immediate Transcriptional Burst'. *Frontiers in Cellular and Infection Microbiology* 9, [online]. Available at: <https://www.frontiersin.org/articles/10.3389/fcimb.2019.00272/full> [accessed 4 Feb 2021].
- SCHUMACHER, Maria A. et al. 2002. 'Structural Basis for Cooperative DNA Binding by Two Dimers of the Multidrug-Binding Protein QacR'. *The EMBO Journal* 21(5), [online], 1210–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC125875/> [accessed 7 Sep 2021].
- SEDLYAROVA, Nadezda et al. 2016. 'SRNA-Mediated Control of Transcription Termination in E. Coli'. *Cell* 167(1), [online], 111-121.e13. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5040353/> [accessed 25 Mar 2022].
- SENARATNE, Ryan H. et al. 2008. '*Mycobacterium tuberculosis* Strains Disrupted in Mce3 and Mce4 Operons Are Attenuated in Mice'. *Journal of Medical Microbiology* 57(2), [online], 164–70. Available at: <https://www.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.47454-0> [accessed 16 Sep 2021].

- SESHASAYEE, Aswin S. N., Gillian M. FRASER, M. Madan BABU and Nicholas M. LUSCOMBE. 2009. 'Principles of Transcriptional Regulation and Evolution of the Metabolic System in E. Coli'. *Genome Research* 19(1), [online], 79–91. Available at: <https://genome.cshlp.org/content/19/1/79> [accessed 2 Sep 2021].
- SHARMA, Cynthia M. et al. 2010. 'The Primary Transcriptome of the Major Human Pathogen Helicobacter Pylori'. *Nature* 464(7286), [online], 250–5. Available at: <https://www.nature.com/articles/nature08756> [accessed 6 Sep 2021].
- SHECHTMAN, Orit. 2013. 'The Coefficient of Variation as an Index of Measurement Reliability'. In Suhail A. R. DOI and Gail M. WILLIAMS (eds.). *Methods of Clinical Epidemiology*. Berlin, Heidelberg: Springer, 39–49. Available at: https://doi.org/10.1007/978-3-642-37131-8_4 [accessed 4 Feb 2023].
- SHIMONO, Nobuyuki et al. 2003. 'Hypervirulent Mutant of *Mycobacterium tuberculosis* Resulting from Disruption of the Mce1 Operon'. *Proceedings of the National Academy of Sciences* 100(26), [online], 15918–23. Available at: <https://www.pnas.org/content/100/26/15918> [accessed 16 Sep 2021].
- SIDDIQI, Noman et al. 2002. 'Molecular Characterization of Multidrug-Resistant Isolates of *Mycobacterium tuberculosis* from Patients in North India'. *Antimicrobial Agents and Chemotherapy* 46(2), [online], 443–50. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC127030/> [accessed 16 Feb 2016].
- SILVA, Pedro E. A. et al. 2001. 'Characterization of P55, a Multidrug Efflux Pump In *Mycobacterium Bovis* and *Mycobacterium tuberculosis*'. *Antimicrobial Agents and Chemotherapy* 45(3), [online], 800–4. Available at: <https://journals.asm.org/doi/10.1128/AAC.45.3.800-804.2001> [accessed 7 Sep 2021].
- SINGH, Amit et al. 2003. 'MymA Operon of *Mycobacterium tuberculosis*: Its Regulation and Importance in the Cell Envelope'. *FEMS Microbiology Letters* 227(1), [online], 53–63. Available at: [https://doi.org/10.1016/S0378-1097\(03\)00648-7](https://doi.org/10.1016/S0378-1097(03)00648-7) [accessed 9 Sep 2021].
- SIU, Gilman Kit Hang, Wing Cheong YAM, Ying ZHANG and Richard Y. T. KAO. 2014. 'An Upstream Truncation of the FurA-KatG Operon Confers High-Level Isoniazid Resistance in a *Mycobacterium tuberculosis* Clinical Isolate with No Known Resistance-Associated Mutations'. *Antimicrobial Agents and Chemotherapy* 58(10), [online], 6093–100. Available at:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4187958/> [accessed 7 Sep 2021].

- SLAGER, Jelle, Rieza APRIANTO and Jan-Willem VEENING. 2018. 'Deep Genome Annotation of the Opportunistic Human Pathogen *Streptococcus Pneumoniae* D39'. *Nucleic Acids Research* 46(19), [online], 9971–89. Available at: <https://doi.org/10.1093/nar/gky725> [accessed 6 Sep 2021].
- SMYTH, G. K. 2005. 'Limma: Linear Models for Microarray Data'. In Robert GENTLEMAN et al. (eds.). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York, NY: Springer, 397–420. Available at: https://doi.org/10.1007/0-387-29362-0_23 [accessed 17 Jun 2022].
- SPEERS, David J. 2006. 'Clinical Applications of Molecular Biology for Infectious Diseases'. *Clinical Biochemist Reviews* 27(1), [online], 39–51. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1390794/> [accessed 16 Sep 2021].
- SPIETH, John et al. 1993. 'Operons in *C. Elegans*: Polycistronic MRNA Precursors Are Processed by Trans-Splicing of SL2 to Downstream Coding Regions'. *Cell* 73(3), [online], 521–32. Available at: <http://www.sciencedirect.com/science/article/pii/009286749390139H> [accessed 26 Jan 2021].
- STACEY, Sean D., Danielle A. WILLIAMS and Christopher L. PRITCHETT. 2017. 'The *Pseudomonas Aeruginosa* Two-Component Regulator AlgR Directly Activates RsmA Expression in a Phosphorylation-Independent Manner'. *Journal of Bacteriology* 199(18), [online], e00048-17. Available at: <https://journals.asm.org/doi/full/10.1128/JB.00048-17> [accessed 14 Sep 2021].
- STEWART, Graham R. et al. 2002. 'Dissection of the Heat-Shock Response in *Mycobacterium tuberculosis* Using Mutants and Microarrays: A List of the 100 ORFs Most Highly Induced by Heat Shock Is Provided as Supplementary Data with the Online Version of This Paper ([Http://Mic.Sgmjournals.Org](http://Mic.Sgmjournals.Org)).' *Microbiology*, 148(10), [online], 3129–38. Available at: <https://www.microbiologyresearch.org/content/journal/micro/10.1099/00221287-148-10-3129> [accessed 10 Nov 2020].
- SURESH, Naga et al. 2006. 'RpoB Gene Sequencing and Spoligotyping of Multidrug-Resistant *Mycobacterium tuberculosis* Isolates from India'. *Infection, Genetics and Evolution* 6(6), [online], 474–83. Available at: <http://www.sciencedirect.com/science/article/pii/S1567134806000359> [accessed 16 Feb 2016].

- TABOADA, Blanca, Ricardo CIRIA, Cristian E. MARTINEZ-GUERRERO and Enrique MERINO. 2012. 'ProOpDB: Prokaryotic Operon DataBase'. *Nucleic Acids Research* 40(D1), [online], D627–31. Available at: <https://doi.org/10.1093/nar/gkr1020> [accessed 6 Sep 2021].
- TABOADA, Blanca, Karel ESTRADA, Ricardo CIRIA and Enrique MERINO. 2018. 'Operon-Mapper: A Web Server for Precise Operon Identification in Bacterial and Archaeal Genomes'. *Bioinformatics* 34(23), [online], 4118–20. Available at: <https://academic.oup.com/bioinformatics/article/34/23/4118/5040321> [accessed 11 Dec 2019].
- TABOADA, Blanca, Cristina VERDE and Enrique MERINO. 2010. 'High Accuracy Operon Prediction Method Based on STRING Database Scores'. *Nucleic Acids Research* 38(12), [online], e130–e130. Available at: <https://doi.org/10.1093/nar/gkq254> [accessed 6 Sep 2021].
- TALWAR, Sakshi et al. 2020. 'Role of VapBC12 Toxin-Antitoxin Locus in Cholesterol-Induced Mycobacterial Persistence'. Edited by Theodore M. Flynn. *mSystems* 5(6), [online], e00855-20. Available at: <https://journals.asm.org/doi/10.1128/mSystems.00855-20> [accessed 7 Jul 2022].
- TELENTI, Amalio et al. 1997. 'The Emb Operon, a Gene Cluster of *Mycobacterium tuberculosis* Involved in Resistance to Ethambutol'. *Nature Medicine* 3(5), [online], 567–70. Available at: <https://www.nature.com/articles/nm0597-567> [accessed 8 Sep 2021].
- TELENTI, Dr Amalio and Michael ISEMAN. 2012. 'Drug-Resistant Tuberculosis'. *Drugs* 59(2), [online], 171–9. Available at: <http://link.springer.com/article/10.2165/00003495-200059020-00002> [accessed 4 Feb 2016].
- THANASSI, D. G., L. W. CHENG and H. NIKAIDO. 1997. 'Active Efflux of Bile Salts by Escherichia Coli'. *Journal of Bacteriology* 179(8), 2512–8.
- THERNEAU, TM and EJ ATKINSON. 2022. 'An Introduction to Recursive Partitioning Using the RPART Routines'. Available at: <https://mran.microsoft.com/web/packages/rpart/rpart.pdf>.
- TJADEN, Brian et al. 2002. 'Transcriptome Analysis of Escherichia Coli Using High-Density Oligonucleotide Probe Arrays'. *Nucleic Acids Research* 30(17), [online], 3732–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC137427/> [accessed 4 Sep 2021].

- TJADEN, Brian. 2019. 'A Computational System for Identifying Operons Based on RNA-Seq Data'. *Methods* [online]. Available at: <http://www.sciencedirect.com/science/article/pii/S1046202318303426> [accessed 23 Oct 2019].
- TJADEN, Brian. 2020a. 'A Computational System for Identifying Operons Based on RNA-Seq Data'. *Methods* 176, 62–70.
- TJADEN, Brian. 2020b. 'A Computational System for Identifying Operons Based on RNA-Seq Data'. *Methods* 176, [online], 62–70. Available at: <https://www.sciencedirect.com/science/article/pii/S1046202318303426> [accessed 6 Sep 2021].
- VOSS, Martin, Manfred NIMTZ and Silke LEIMKÜHLER. 2011. 'Elucidation of the Dual Role of Mycobacterial MoeZR in Molybdenum Cofactor Biosynthesis and Cysteine Biosynthesis'. *PLOS ONE* 6(11), [online], e28170. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028170> [accessed 1 Oct 2021].
- WALTERS, D. M., R. RUSS, H. -J. KNACKMUSS and P. E. ROUVIÈRE. 2001. 'High-Density Sampling of a Bacterial Operon Using mRNA Differential Display'. *Gene* 273(2), [online], 305–15. Available at: <https://www.sciencedirect.com/science/article/pii/S0378111901005972> [accessed 17 Sep 2021].
- WANG, Ligu, Shengqin WANG and Wei LI. 2012a. 'RSeQC: Quality Control of RNA-Seq Experiments'. *Bioinformatics* 28(16), [online], 2184–5. Available at: <https://doi.org/10.1093/bioinformatics/bts356> [accessed 22 Jun 2022].
- WANG, Ligu, Shengqin WANG and Wei LI. 2012b. 'RSeQC: Quality Control of RNA-Seq Experiments'. *Bioinformatics* 28(16), [online], 2184–5. Available at: <https://doi.org/10.1093/bioinformatics/bts356> [accessed 23 Mar 2022].
- WANG, Shuqin et al. 2007. 'A Multi-Approaches-Guided Genetic Algorithm with Application to Operon Prediction'. *Artificial Intelligence in Medicine* 41(2), [online], 151–9. Available at: <https://www.sciencedirect.com/science/article/pii/S0933365707000966> [accessed 3 Sep 2021].
- WANG, Xiaoyu et al. 2018. 'Mycobacterium tuberculosis Toxin Rv2872 Is an RNase Involved in Vancomycin Stress Response and Biofilm Development'. *Applied Microbiology and Biotechnology* 102(16),

[online], 7123–33. Available at: <https://doi.org/10.1007/s00253-018-9132-0> [accessed 7 Jun 2022].

WANG, Zhong, Mark GERSTEIN and Michael SNYDER. 2009. ‘RNA-Seq: A Revolutionary Tool for Transcriptomics’. *Nature reviews. Genetics* 10(1), [online], 57–63. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/> [accessed 13 Sep 2021].

WATTAM, Alice R. et al. 2014. ‘PATRIC, the Bacterial Bioinformatics Database and Analysis Resource’. *Nucleic Acids Research* 42(Database issue), D581-591.

WEBBER, Mark A. et al. 2009. ‘The Global Consequence of Disruption of the AcrAB-TolC Efflux Pump in *Salmonella* Enterica Includes Reduced Expression of SPI-1 and Other Attributes Required To Infect the Host’. *Journal of Bacteriology* 191(13), [online], 4276–85. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698494/> [accessed 7 Sep 2021].

WELLS, Ryan M. et al. 2013. ‘Discovery of a Siderophore Export System Essential for Virulence of *Mycobacterium tuberculosis*’. *PLoS Pathogens* 9(1), [online], e1003120. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3561183/> [accessed 7 Dec 2021].

DE WELZEN, Lynne et al. 2017. ‘Whole-Transcriptome and -Genome Analysis of Extensively Drug-Resistant *Mycobacterium tuberculosis* Clinical Isolates Identifies Downregulation of EthA as a Mechanism of Ethionamide Resistance’. *Antimicrobial Agents and Chemotherapy* 61(12), e01461-17.

WERNGREN, Jim, Erik ALM and Mikael MANSJÖ. 2017. ‘Non-PncA Gene-Mutated but Pyrazinamide-Resistant *Mycobacterium tuberculosis*: Why Is That?’ *Journal of Clinical Microbiology* 55(6), [online], 1920–7. Available at: <https://journals.asm.org/doi/full/10.1128/JCM.02532-16> [accessed 14 Sep 2021].

WHO. 2020. ‘WHO | Global Tuberculosis Report’. [online]. Available at: <https://www.who.int/news-room/fact-sheets/detail/tuberculosis> [accessed 2 Sep 2021].

WHO. 2021. ‘Global Shortage of Innovative Antibiotics Fuels Emergence and Spread of Drug-Resistance’. [online]. Available at: <https://www.who.int/news/item/15-04-2021-global-shortage-of->

innovative-antibiotics-fuels-emergence-and-spread-of-drug-resistance
[accessed 8 Oct 2021].

WILLIAMS, Graham. 2011. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. Springer Science & Business Media.

WILSON, Michael et al. 1999. 'Exploring Drug-Induced Alterations in Gene Expression in *Mycobacterium tuberculosis* by Microarray Hybridization'. *Proceedings of the National Academy of Sciences of the United States of America* 96(22), [online], 12833–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC23119/> [accessed 7 Jun 2022].

WIPPERMAN, Matthew F., Nicole S. SAMPSON and Suzanne THOMAS T. 2014. 'Pathogen 'Roid Rage: Cholesterol Utilization by *Mycobacterium tuberculosis*'. *Critical reviews in biochemistry and molecular biology* 49(4), [online], 269–93. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4255906/> [accessed 11 Oct 2021].

WOOLLEY, Robin C. et al. 2005. 'Characterization of the Vibrio Cholerae VceCAB Multiple-Drug Resistance Efflux Operon in Escherichia Coli'. *Journal of Bacteriology* 187(15), [online], 5500–3. Available at: <https://journals.asm.org/doi/full/10.1128/JB.187.15.5500-5503.2005> [accessed 7 Sep 2021].

WORLD HEALTH ORGANIZATION. 2019. *Global Tuberculosis Report 2019*. Available at: <https://apps.who.int/iris/bitstream/handle/10665/329368/9789241565714-eng.pdf?ua=1>.

WORLD HEALTH ORGANIZATION. 2022. 'Global Tuberculosis Report 2022'. [online]. Available at: <https://www.who.int/publications/i/item/9789240061729> [accessed 22 Feb 2023].

VAN DER WOUDE, Marjan W. and Andreas J. BÄUMLER. 2004. 'Phase and Antigenic Variation in Bacteria'. *Clinical Microbiology Reviews* 17(3), [online], 581–611. Available at: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC452554/> [accessed 16 Feb 2016].

XIONG, A. et al. 2000. 'The EmrR Protein Represses the Escherichia Coli EmrRAB Multidrug Resistance Operon by Directly Binding to Its Promoter Region'. *Antimicrobial Agents and Chemotherapy* 44(10),

[online], 2905–7. Available at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC90178/> [accessed 7 Sep 2021].

XU, Zhihong et al. 2018. ‘Transcriptional Approach for Decoding the Mechanism of RpoC Compensatory Mutations for the Fitness Cost in Rifampicin-Resistant *Mycobacterium tuberculosis*’. *Frontiers in Microbiology* 9, [online], 2895. Available at:
<https://www.frontiersin.org/article/10.3389/fmicb.2018.02895> [accessed 16 Sep 2021].

YADA, T., M. NAKAO, Y. TOTOKI and K. NAKAI. 1999. ‘Modeling and Predicting Transcriptional Units of Escherichia Coli Genes Using Hidden Markov Models’. *Bioinformatics (Oxford, England)* 15(12), 987–93.

YANG, Samuel and Richard E ROTHMAN. 2004. ‘PCR-Based Diagnostics for Infectious Diseases: Uses, Limitations, and Future Applications in Acute-Care Settings’. *The Lancet. Infectious Diseases* 4(6), [online], 337–48. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7106425/> [accessed 16 Sep 2021].

YOON, Sung Ho et al. 2011. ‘Parallel Evolution of Transcriptome Architecture during Genome Reorganization’. *Genome Research* 21(11), [online], 1892–904. Available at:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3205574/> [accessed 25 Jan 2021].

ZAIDI, Syed Shujaat Ali and Xuegong ZHANG. 2017. ‘Computational Operon Prediction in Whole-Genomes and Metagenomes’. *Briefings in Functional Genomics* 16(4), [online], 181–93. Available at:
<https://academic.oup.com/bfg/article/16/4/181/2555398> [accessed 18 Apr 2020].

ZHANG, Guo-qing et al. 2006. ‘Operon Prediction Based on SVM’. *Computational Biology and Chemistry* 30(3), 233–40.

ZHANG, Yong-Mei and Charles O. ROCK. 2008. ‘Membrane Lipid Homeostasis in Bacteria’. *Nature Reviews Microbiology* 6(3), [online], 222–33. Available at: <https://www.nature.com/articles/nrmicro1839> [accessed 27 Jul 2022].

ZHANG, Yong-Mei and Charles O. ROCK. 2010. ‘A Rainbow Coalition of Lipid Transcriptional Regulators’. *Molecular microbiology* 78(1), [online], 5–8. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2967205/> [accessed 27 Jul 2022].

- ZHAO, Shanrong, Kurt PRENGER and Lance SMITH. 2013. 'Stormbow: A Cloud-Based Tool for Reads Mapping and Expression Quantification in Large-Scale RNA-Seq Studies'. *ISRN Bioinformatics* 2013, [online], 481545. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4393068/> [accessed 13 Sep 2021].
- ZHENG, Yu et al. 2002. 'Computational Identification of Operons in Microbial Genomes'. *Genome Research* 12(8), 1221–30.
- ZHU, Jun-Hao et al. 2018. 'Rifampicin Can Induce Antibiotic Tolerance in Mycobacteria via Paradoxical Changes in RpoB Transcription'. *Nature Communications* 9(1), [online], 4218. Available at: <https://www.nature.com/articles/s41467-018-06667-3/> [accessed 8 Sep 2021].
- ZWADYK, P, J A DOWN, N MYERS and M S DEY. 1994. 'Rendering of Mycobacteria Safe for Molecular Diagnostic Studies and Development of a Lysis Method for Strand Displacement Amplification and PCR'. *Journal of Clinical Microbiology* 32(9), [online], 2140–6. Available at: <https://journals.asm.org/doi/abs/10.1128/jcm.32.9.2140-2146.1994> [accessed 28 Feb 2023].

