UNIVERSITY OF CALIFORNIA, SAN DIEGO

Bioinformatic Functional Characterization of the Prokaryotic FUPAs through Co-localization and Co-regulatory Analysis: Taking the FU out of the FUPAs

A thesis submitted in partial satisfaction of the requirements for the degree

Master of Science

in

Biology

by

Russell Wade De La Mare

Committee in charge:

Milton H. Saier, Jr., Chair
Kathleen French
Eric Allen

2012

The Thesis of Russell Wade De La Mare is approved and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____
Chair

University of California, San Diego

2012

DEDICATION

I dedicate this thesis to my family, friends, Dr. Saier and all of the members of Saier Lab with whom I have had the pleasure of working and those whose work made this project possible. This thesis is also dedicated to all of the professors who have prepared me for this and future undertakings, especially Dr. Kathy French, for providing much of the mentorship I sorely needed. A huge “thank you” is owed to Rostislav Castillo, Dorjee Tamang, Andrei Osterman, Dmitry Rodionov and Pavel Novichkov for their support and guidance.

Naturally, I would like to thank Dr. Milton H. Saier, Jr., for his guidance, support and friendship, as well as his passion and enthusiasm for scientific research, which he has inspired in many students, me included. In the four years that I have known Dr. Saier, I have learned a lot about science and life in general. From times when we were alone in the lab to those when it was so crowded one couldn’t hear themselves think to that time when no one was allowed in at all for all the wrong reasons, he has been a fine friend and a mentor, as well as inspiring us to work hard by setting the example. As Dr. Saier once told me, he made it his responsibility to “raise the kids in the lab”. Dr. Saier’s work beyond academia has also inspired me and expanded my worldview in ways I am glad I will never part from. I sincerely thank Dr. Saier and all of the members of Saier Lab.

EPIGRAPH

Insert clever quote here

Anonymous

TABLE OF CONTENTS

Signature Page ……...…… iii

Dedication ……..…… iv

Epigraph …… v

Table of Contents ……...…… vi

Acknowledgements …… viii

Abstract of the Thesis …… ix

Introduction …… 1

Methods ……. 4

FUPA23 ATPase Family ……..…….....…… 6

FUPA24 ATPase Family ……...…… 32

FUPA25 ATPase Family ……..…… 59

FUPA26 ATPase Family ……..…… 103

FUPA27 ATPase Family ……..…… 113

FUPA28 ATPase Family ……..……..…… 166

FUPA29 ATPase Family …… 169

FUPA30 ATPase Family ……..…… 205

FUPA31 ATPase Family ……..…… 243

FUPA32 ATPase Family ……...……..…… 253

Discussion…...…….. 303

Table 1: Summary of functional predictions made for FUPA23-32 …… 310

References ……..…… 311

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Milton H. Saier, Jr., for his support and for serving as the chair of my committee. Similarly, I wish to acknowledge Dr. Kathleen French and Dr. Eric Allen for taking the time and making the effort to serve on my committee. Having enjoyed taking one engaging course and assisting with teaching thrice more with Dr. French, it is an honor to have her evaluate my performance once again, in the context of a thesis defense. Dr. Allen has been especially generous and accommodating in a pinch with his time and advice.

Aside from Dr. Saier, Rostislav Castillo and Andrei Osterman were my primary sources of mentorship and were invaluable in helping me get over the learning curve for the research performed in our lab. I am grateful to both them both and they most certainly deserve honorary mentions.

I would like to acknowledge Dorjee Tamang, Robert Olson, Dmitry Rodionov and Pavel Novichkov for all of their help on the technical side of this project. I thank Andrew Lukosus for all of his help on the administrative side and Mark Whelan for his work as a TA coordinator, which allowed me to experience teaching a myriad of lecture courses.

ABSTRACT OF THE THESIS

Bioinformatic Functional Characterization of the Prokaryotic FUPAs through Co-localization and Co-regulatory Analysis: Taking the FU out of the FUPAs

by

Russell Wade De La Mare

Master of Science in Biology

University of California, San Diego, 2012

Professor Milton H. Saier, Jr., Chair

P-type ATPases are ubiquitous in all domains of life. Chan, et al. (2010) analyzed P-type ATPases in all the major prokaryotic phyla for which complete genome sequence data were available. The P-type ATPase superfamily consists of thirty-two recognized families, 17 of which are strictly found in prokaryotes. The first seven of these are Families 1-7, which are generally well characterized. Ten functionally uncharacterized

P-type ATPase (FUPA) families were identified in as well, FUPA23 through FUPA32. Here, these families are analyzed individually rather than by phyla using the genomic context program SEED (Overbeek et al., 2005) and the co-regulation prediction program RegPredict (Novichkov PS, et al., 2010). By examining the known function of genes co-localized near those encoding FUPAs and using co-localization as a proxy for co-regulation, and finding other genes regulated by similar regulatory sequences, we are able to make highly robust predictions about the putative functions of these ATPases and functionally characterize them. In the process of doing this, unique characteristics of the proteins are illuminated and discussed in the context of phylogeny, environmental context and horizontal gene transfer and how each contributes to the functionality of these ATPases in genomes in which they are found.

Introduction

P-type ATPases are found in all domains of life. According to the Transporter Classification Database (http://www.tcdb.org), there are 32 families of P-type ATPases. Nine families are characterized, all as cation transporters, except for Family 8, found only in eukaryotes, which transports phospholipids from the outer leaflet to the inner leaflet of the membrane. Families 23-32 are functionally uncharacterized P-type ATPases, or FUPAs. The use of functional, phylogenetic, and membrane topology information extracted from over 10,000 publications on functional data and novel transport systems has allowed our lab to classify over 5,000 transport proteins into over 600 families. The fruits of this work can be found in the IUBMB approved Transporter Classification Database (http://www.tcdb.org), a curated database which employs the TC system, and which is analogous to the function-only based Enzyme Commission (EC) system (http://www.chem.qmul.ac.uk/iubmb/enzyme/; Saier et al. 2005; 2009). This research emphasizes the synergy of database collaboration as the SEED database (http://pubseed.theseed.org/) and TCDB have been used to augment the content found in both. My study focuses on the functionally uncharacterized P-type ATPases (FUPA) subfamilies (TCID 3.A.3.23-32) of transmembrane proteins such that the gaps of knowledge for these protein subfamilies are bridged, leading to more directed future experiments, and improved resolution and clarity regarding the functions of the FUPAs, so that SEED and TCDB can be updated and expanded.

Transport systems play a crucial role in every process of life. Some examples include nutrient transport, metabolite excretion, essential cation acquisition and

allocation, drug/toxin secretion, establishing electrochemical gradients, macromolecular export, stress response-related transport and transport of signaling molecules (Busch et al. 2002). These functions have all been suggested, to varying degrees, for various P-type ATPases, necessitating consideration for any evidence of each.

Stress response-related transport is of special interest here, as many of the FUPA proteins have been implicated in this type of transport (Chan, et al., 2010). UspA expression rises when the organism is exposed to stress conditions. UspA enhances cell survival during prolonged exposure these conditions. The occurrence of multiple universal stress protein-encoding genes amongst many of those genes that encode FUPAs further supports the hypothesis that many of the FUPAs participate in stress-condition response and endurance.

The functionally uncharacterized P-type ATPases (FUPA) subfamilies (TCID 3.A.3.23-32), are comprised of over 1000 members spanning the prokaryotic and archaeal domains. Within the prokaryotic domain, these proteins have been identified in both Gram-positive and Gram-negative bacteria. Various genomic sources are found for most phylogenetic clusters, which suggests horizontal gene transfer (HGT) is responsible for some orthologues (Yen, et al., 2009). The majority of prokaryotic protein members range in size from 600 to 1650 amino acids with few exceptions. The archaeal members are similar in size to the prokaryotic members, except that due to the absence of the fusion-protein family, FUPA24 (Chan, et al., 2010), the upper limit in size is 1046 amino acids.

The functions of nearly all of the FUPA family members have not been assigned and cannot be assumed due to the fact that relatively few amino acids determine the substrates of a given substrate, and not all determining amino acid motifs are known (Chan, et al., 2010).

The central purpose of this work is to suggest related functions for members of the FUPA subfamilies. It is extremely unlikely that function characterization of the FUPA subfamilies will result in direct biomedical applications. However, complete characterization of the FUPA subfamilies, and transporter families as a whole, as well as a deeper understanding of proteomes, are useful in the long term in myriad ways as they help in the discovery of antimicrobial drug therapies, industrial microbe applications and artificial probiotics.

To accomplish these ends, several bioinformatic tools are applied. In characterizing “hypothetical” or “unknown” proteins encoded near the FUPAs, which are completely uninformative regarding potential shared function, any possible functional domains present in the hypothetical proteins will be identified using the Conserved Domain Database (http://www.ncbi.nlm.nih.gov/cdd) WHAT, TMHMM 2.0 and HMMTOP, which allows for functional determinations such as a more accurate description of the topology of the FUPAs and their surrounding proteins. Genome context analysis using SEED (http://www.theseed.org), and transcription factor binding site analyses, using RegPredict (http://regpredict.lbl.gov/regpredict/) will be performed in order to predict possible related functions (Overbeek et al., 2005; Novichkov et al., 2010a; 2010b).

Methods

The BLAST function of TCDB was used in conjunction with the PsiBLAST function of NCBI to identify P-type ATPase family members in over 750 prokaryotic organisms in Chan, et al., 2010. Approximately 94 FUPAs falling into families 23-32 were identified. A continuation of the methods described in Chan, et al., 2010 were used to identify and incorporate twenty-seven additional organisms here, based on new developments in TCDB as well as points of interest found through SEED analysis.

Predicting possible related functions was a multi-step, often cyclic process beginning with genome context analyses performed using The SEED-Viewer and Subsystem Editor, which provided for the review of over 4,500 curated genomes, as well as the capacity to create FUPA subfamily-specific subsystems, in order to discover homologous genes, their operon context, and consequently their known or putative roles in other organisms (Overbeek, et al., 2005). Many of the genes found adjacent to the FUPA genes were also uncharacterized or poorly curated.

RegPrecise and RegPredict, which provide strong predictions for the identification of transcription factor binding sites (Novichkov et al., 2010a; 2010b), were also used to expand the base of prediction sources available and attempt to catch as much of each FUPA's regulon as possible to incorporate into functional predictions. Many of the genes uncovered here required characterization as well.

Previously uncharacterized proteins only annotated as hypothetical proteins, as well as other poorly curated genes, in SEED and RegPredict were identified as thoroughly as they could be using PSI-BLAST. Genes were identified as good

candidates for further characterization based on their proximity to FUPA genes and their frequency of co-localization, as well as their inclusion in RegPredict predictions for co-regulation. The first iteration of the PSI-BLAST operation (NCBI) was performed with hypothetical proteins using normal settings, with the output set to 500 sequences, and with a cutoff of e^{-4}. A second iteration was performed in the same manner, but to minimize false positives, a stricter cutoff of e^{-6} was used (Altschul et al., 1990; 1997). Topological analyses of single protein sequences were performed using WHAT, TMHMM 2.0, and HMMTOP (Zhai and Saier, 2001a; Tusnády, et al., 1998; 2001; Bailey et al., 1994; 1998). A search for functional domains within all BLASTed proteins was performed using the conserved domain database (CDD) of NCBI (Marchler-Bauer et al., 2009).

Whenever an EC number was available for a protein encoded adjacent to a FUPA gene or predicted to share a regulatory sequence with one, the number was used to find known metal cation and other ion requirement data for that EC family using BRENDA (http://www.brenda-enzymes.org/) (Scheer M, et al., 2011).

The FUPA23 ATPase Family

FUPA23 is a family of Type II P-type ATPases with 8 members of sizes ranging from 790-837 aas in length and a topology of 10 TMS, with 1 & 2 pairing and found around 80 aas, 3 & 4 pairing around 200-300 aas and 5-10 clustered in the last 200 aas of the protein. The family's nearest hit in TCDB is Family 2 (3.A.3.2.-), then Family 3 (3.A.3.3.-); FUPA24 (3.A.3.24.-) sometimes comes up before them though, presumably because of its Type II C-terminal portion. FUPA23 (TC 3.A.3.23) and FUPA24 (3.A.3.24) are known to cluster loosely together with Family 2 (Type II) Ca^{2+} -ATPases [Chan, et al., 2010]. Two paralogs are found in *S.coelicolor*; all other members are distributed one per organism. As described in Chan et al. (2010), the relationships between these proteins reflect the respective organisms' orthology, and assuming that Sco7 and Sco11 arose by gene duplication before the divergence of *S. coelicolor* from *S. avermitilis*.

Thirteen organisms are described in Chan, et al. (2010); two with 2 FUPA23s, for 15 total. No novel FUPA23s were found in SEED, so this analysis is restricted to those original 15.

The gene for **Blo5** (*Bifidobacterium longum* NCC2705 837aa, gi:23465956) is the fourth-to-last transcribed gene in a gene cluster of 13 genes. Due to distances of greater than 50bp between several of these genes, they are not suggested to be contained within a single operon. However, their uniform direction of transcription and their conserved proximity to each other does raise the question. The first gene in the cluster in

the direction of the FUPA23 is included due to its relation to the subsequent genes, but it is 301bp upstream of the next (2). Starting from the beginning, this series is as follows:

1) dipeptide ABC transporter, periplasmic substrate-binding component DppA2 (TC 3.A.1.5.2) (Zn^{2+} dependent)
2) 301bp later, dipeptide transport system permease protein DppB (TC 3.A.1.5.2)
3) 21bp later, dipeptide transport system permease protein DppC (TC 3.A.1.5.2)
4) 23bp later, dipeptide transport ATP-binding protein DppD (TC 3.A.1.5.2)
5) 105bp later, exodeoxyribonuclease III (EC 3.1.11.2) (Mg^{2+} requiring, Mn^{2+} and Ca^{2+} can substitute [Blair JM et al., 1969]). *Note: in the strains *B.longum subsp. infantis* ATCC 15697 and *B.longum* DJO10A, this gene is transcribed in the opposite direction, with 125bp between it and gene (6)
6) 68bp later is a DUF3710 superfamily protein-encoding gene.
7) 16bp later is a gene for a 252aa protein known as "PROBABLE CONSERVED INTEGRAL MEMBRANE ALANINE AND LEUCINE RICH PROTEIN" with 99% identity to "ABC-type Mn^{2+}/Zn^{2+} transport systems, permease components" in NCBI pBLAST
8) -49bp later (overlap), SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase
9) 12bp later, ABC-type sugar transport system, periplasmic component
10) 91bp later, FUPA23 P-type ATPase
11) 120 bp later, aconitate hydratase (EC 4.2.1.3) (binds 1 4Fe-4S cluster per subunit [Tang Y, et al., 2005] and possibly Mg^{2+} [Tsuchiya D, et al., 2008]) *This gene

and a homolog of “SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase” are also found with the FUPA23 of [*Renibacterium salmoninarum* ATCC 33209], a gram positive salmonid parasite

12) 146bp later, DNA-damage-inducible protein J
13) 0bp later, PilT domain-containing protein

RegPredict suggests that the genes for FUPA23 through PilT compose an operon, with a palindromic regulatory sequence (palindrome is in caps) of cCTgAatcATcccTgAGc at a score of 6.85. It also predicts a SAM-dependent methyltransferase type 11 and “Uncharacterized conserved protein, contains double-stranded beta-helix domain” to be regulated by the related regulatory sequence of cCTgAatcgtcccTgAGc at a score of 6.39.

Lxy2 (*Leifsonia xyli subsp. xyli* str. CTCB07 793aa 5 gi:0954725) is encoded 4th in a series of co-directional genes that begins and proceeds as follows:

1) Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (Mg^{2+} is a cofactor, Mn^{2+}, Co^{2+} and Zn^{2+} can substitute [Mustafi, D et al., 2003])
2) 138bp later, DUF3710 superfamily protein
3) 9bp later, a gene for a 252aa protein known as “PROBABLE CONSERVED INTEGRAL MEMBRANE ALANINE AND LEUCINE RICH PROTEIN”
4) 47bp later, FUPA23 P-type ATPase
5) 101bp later, 2-methylisocitrate dehydratase (EC 4.2.1.99)/Aconitate hydratase (EC 4.2.1.3) (see Blo5 above, #11)

6) 78bp later, 1-deoxy-D-xylulose 5-phosphate synthase (EC 2.2.1.7) (divalent cation Ligand required; Mg^{2+}, Mn^{2+} work best, Zn^{2+} to a lesser extent [Bailey AM, et al., 2002])

Sharing a regulatory sequence, 41bp away and divergently transcribed from "deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23)" is a single gene encoding ~150-200aa Ala rich Membrane protein.

R.salmoninarum shares the FUPA23, aconitate hydratase (EC 4.2.1.3), as well as possessing a putative transporter/amino acid permease (no good hits in TCDB but in NCBI:amino acid/polyamine/organocation transporter APC superfamily in *A.phenanthrenivorans* came up with a very good score) and deoxyribonuclease/rho motif-related TRAM both preceding the ATPase co-directionally but with too much distance between to be sure of coregulation. *Sanguibacter keddiei keddieii* has a similar arrangement of genes to *R.salmoninarum*, but is lacking the aconitate hydratase. Finally, *Beutenbergia caverna cavernae* also possesses the arrangement of *S.keddieii*, but the 4 genes are 7bp overlapping, 3bp overlapping and 77bp later.

RegPredict offers little in the way of coregulation predictions. The predicted regulatory sequence preceding the FUPA23 is CggCCgaCGcgGGaaG (score=6.19). The only genes predicted to be coregulated are a 2-cistron operon starting with *gntR*, a transcriptional regulator of the GntR family followed by *adh2*, an alcohol dehydrogenase (locus tags Lxx23620 adn Lxx23600 respectively). These have a regulatory sequence preceding them of CgcCcgcagcgaGaaG (score=5.24).

The genes for *Mycobacterium avium* and *M.bovis* FUPA23s **Mav1** (*M. avium subsp. paratuberculosis* K-10 790aa, gi:41406941) & **Mbo5** (*M. bovis* AF2122/97 797aa, gi:31792096) are found in similar operons/conserved series which begin and proceed as follows:

1) enoyl-CoA hydratase (EC 4.2.1.17) (Fe^{2+} in *Clostridium aminobutyricum*, each 56 kDa subunit of the homotetrameric enzyme contains one FAD and a $[Fe_4S_4]^{2+}$ cluster, [Friedrich P, et al., 2008) (only reference to metal/ion]
2) 6bp later, Outer membrane protein romA
3) -15bp (overlap) later, Beta-lactamase (EC 3.5.2.6) (metalloenzyme, 2 Zn^{2+} binding sites but divalent Ni^{2+}, Cu^{2+}, Co^{2+}, Cd^{2+}, Mg^{2+} and Mn^{2+} may also work) (30bp later in *M.bovis* AF2122/97 in SEED, but probably a sequencing error)
4) -3bp (overlap) later, FUPA23
5) 63bp later, hypothetical protein, no useful PSI BLAST
6) 4bp later, Polyketide cyclase/dehydrase related
7) 98bp later, putative hydroxylase (Mbo5 only)
8) 69bp later, probable conserved TMB protein

Divergently transcribed, 28bp away, an operon encoding the following proteins is found:

1) Acetyl-coenzyme A carboxyl transferase alpha chain (EC 6.4.1.2)/Acetyl-coenzyme A carboxyl transferase beta chain (EC 6.4.1.2) (the enzyme requires

Mg^{2+} or Mn^{2+} for coordinating the ATP phosphates for catalysis, possibly Zn^{2+} [Benson BK, et al., 2008])

2) This varies between species
    a. 115bp later, C-term 1/2 of pentachlorophenol monooxygenase (**Mav1** only, last for the Mav1 containing series)
    b. 131bp later, Two component transcriptional response regulator protein PrrA (Mbo5s series only)
3) 11bp later, Sensor-type histidine kinase PrrB (EC 2.7.13.3) (Heme/Iron, Mg^{2+} cofactors [Shrivastava R, et al., 2007], Mg^{2+} also may be required [Del Papa MF, et al., 2008]) (**Mbo5** only, last for Mbo5's series)

RegPredict identified a non-TATA box palindromic regulatory sequence for these two operons (described above): CcaGccgtGataCgactCacG with a score of 5.99 in front of the Enoyl-CoA hydratase and acctgggaGataCcgatcatg with a score of 5.54 for Acetyl-coenzyme A carboxyl transferase alpha chain (EC 6.4.1.2)/Acetyl-coenzyme A carboxyl transferase beta chain (EC 6.4.1.2). RegPredict also finds *M.tuberculosis* and *M.bovis* to have arylsulfatase AtsB and esterase lipoprotein LpqC regulated by the site gcgctggGGa|t|aCCgatcatg with a score of 5.44 and a probable acetyl-CoA acetyltransferase FadA6 (EC 2.3.1.9) (activity greatest at 5mM Mg^{2+}, Mn^{2+} or Ca^{2+}) and putative uncharacterized protein BCG_3619c regulated by the site gcgGtGgGGg|t|aCCaCtCact with a score of 5.27.

Due to the overlap of this FUPA23 with Beta-lactamase (EC 3.5.2.6) and likely expression with Acetyl-coenzyme A carboxyl transferase alpha chain (EC

6.4.1.2)/Acetyl-coenzyme A carboxyl transferase beta chain (EC 6.4.1.2), this FUPA23 transporter is predicted to transport Zn^{2+}. There is also a possibility that it transports Mg^{2+} or Mn^{2+}, but these are less likely considering the existing literature regarding the aforementioned enzymes.

Nfa6 (*Nocardia farcinica* IFM 10152 650aa, gi:54023752) is not found encoded with commonly co-localized genes according to SEED. Only a gene for a Phenazine biosynthetic protein of the PhzF family is found transcribed co-directionally, 58bp after it. It is tail-to-tail with a gene for the protein RarD, a 10 TMS possible permease (TC# 2.A.7.7.1, Chloramphenicol-sensitive protein RarD), with 84bp between them. Following *rarD* in the same direction 72 bp later is an ADP-ribosylglycohydrolase gene. While the glycohydrolase is almost certainly not expressed with FUPA23, *rarD* likely is. *RarD* is typically expressed with the genes for ribosomal large subunit pseudouridine synthase D (EC 4.2.1.70, dependent on presence of Mg^{2+}, Co^{2+}, Fe^{2+} or Mn^{2+}, inhibited by Zn^{2+} and Ni^{2+} [Heinrikson RL, et al., 1964; Preumont A, et al., 2008]), which here is found in an operon that ends head-to-head with phenazine biosynthesis protein, DNA polymerase III alpha subunit (EC 2.7.7.7) and a transcriptional regulator in the TetR family. Considering the way RarD and ribosomal large subunit pseudouridine synthase D are situated, it is as if the FUPA23 and Phenazine biosynthesis protein interrupt a normally continuous operon. This is supported by the presence of such an operon in *Rhodococcus jostii* RHA1, which has an operon starting with a gene encoding Lipoprotein Signal peptidase (EC 3.4.23.36) followed sequentially by the gene for

ribosomal large subunit pseudouridine synthase D, a gene for a hypothetical protein particular to *N.farcinica* and *R.jostii* and lacking any indication of function, followed by *rarD*. This is followed by a DNA polymerase III alpha subunit (EC 2.7.7.7) gene in *R.jostii* only. The transcriptional regulator in the TetR family gene is also found preceding DNA PolIII in the Mycobacterial operon, which is lacking the hypothetical protein.

RegPredict supports the presence of an operon starting with the genes for the lipoprotein signal peptidase (EC 3.4.23.36) and continuing through DNA pol III alpha subunit in two species of Rhodococcus: *R.sp.* RHA1 and *R.opacus* B4, with sequences of cCcaggttaGtGCgCgcgggaGc (score=5.02) and cCaggCtaGtG|CgCgcGggaGc (score=4.90) respectively. For *N.farcinica*, two similar sequences are found for the divided former operon, whose genes are found separately in this organism, one preceding the hypothetical protein gene and one preceding *rarD*. They are cccggccgaCg|gGcgctggctc (score=4.88) and GCcagcCtaaG|CccgGtggcGC (score=5.12), also respectively. The FUPA23 (and Phenazine biosynthetic protein in *N.farcinica*) in *N.farcinica* and *R.sp*. RHA1 are suggested to be regulated by similar sequences: CcGaCacgaCCGGcgcgGgCcG (score=5.12) and GCGcgtctGGCGCCgctgtCGC (score=4.92). Finally, *N.farcinica* has two regulatory sites, one major and one minor, regulating a gene for a transcriptional regulator in the TetR family. Their sequences are CCGactcCagccgcGcgggCGG (score=5.07) and gccgCgctgcCGcggcgGgccg (score=4.52), respectively. The major regulatory sequence precedes an integral membrane protein gene (only NCBI name within 10e-5) that overlaps the TetR family

protein gene by 10 bp. The minor sequence is predicted to immediately precede the TetR family protein gene, so it either genuinely lies within the end of the integral membrane protein gene sequence or is a false prediction. Moreover, a TetR family protein gene can be found regulated in parallel with many other genes with regulatory sequences related to that of the FUPA23 ATPase gene, as well as approximately 20 large and small ribosomal subunit proteins.

Sav1 (*Streptomyces avermitilis* MA-4680 797aa, gi:29830442) is encoded with somewhat commonly co-localized genes according to SEED. However, none of these are found in any orientation that implies co-expression. A gene for a 2 TMS, ~89aa putative integral membrane protein which only occurs 4 time, always in Streptomycetes, 3 of these upstream of FUPA23 genes is found in the same orientation as the FUPA23, but it is 157bp upstream from the P-type ATPase and thus may to be co-transcribed. The gene product's closest hit in TCDB is (TC#3.A.1.29.1) UPF0397 protein with and e value of 0.023.

RegPredict predicts a shared regulatory sequence for *S.griseus*, *S.avermitilis* and *S.coelicolor* A3(2) for FUPA23 of tcctcccCtGgtCgGatcttcg with a score of 7.68. The only other predictions using similar regulatory sequences in these organisms are a putative secreted tripeptidyl aminopeptidase gene in *S.griseus* with a regulatory sequence of tcgccccCgGggCgGatcttcc and score of 5.70 and a gene for a transcriptional regulator of the TetR family in *S.avermitilis* with a regulatory sequence of acctccGCtGttCgGCtcttcg and score of 5.69.

The gene encoding **Sco7** (*Streptomyces coelicolor* A3(2) *707aa, gi:21221651) *(*NCBI and SEED describe this protein as having ~796aa)* has only one local co-directional gene somewhat nearby, 745 bp downstream (for a putative transcriptional regulator) and thus not co-transcribed. Divergently, there is an operon 293bp away, which could be co-regulated. This operon starts and proceeds as follows:

1) generic methyltransferase
2) -3bp later (overlap), Anthranilate synthase, aminase component (EC 4.1.3.27), TrpAa
3) -3bp later (overlap), Anthranilate synthase, aminase component (EC 4.1.3.27), TrpAb
4) 17bp later, Anthranilate phosphoribosyltransferase (EC 2.4.2.18) (Mg^{2+} required [Egan AF, et al., 1972])
5) -3bp later (overlap), Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (Ca^{2+}, Mg^{2+}, Mn^{2+} and Na^{+} "significantly affects activity, the optimal concentration is about 0.4-2.0 mM" according to Yang Y, et al. (2006)
6) -3bp later (overlap), 2-keto-3-D-arabino-heptulosonate-7-phosphate synthase II (EC 2.5.1.54), AroA II

RegPredict suggests double regulatory sequences for the FUPA23 of acCtGGTgggtgtggACCtGcc with score 5.58 and atCtggtaCGtcCGggggcGcg with score 4.62 for *S.coelicolor*. *Frankia sp*. EAN1pec and *Frankia sp*. Ccl3 also share the first

sequence for regulation of their FUPA23. A regulatory sequence for the divergent operon of cCgtcGtgcGggCaggCccgGc with a score of 5.3 is also predicted.

The gene for **Sco11** (*S.coelicolor* A3(2) 802aa, gi:21222724) is found only with the gene for a putative integral membrane protein which was mentioned in the analysis of Sav1 which is found in the same orientation as the FUPA23, but it is 100bp upstream from the P-type ATPase and thus not likely to be co-transcribed.

RegPredict analysis of the 100bp intergenic sequence preceding the FUPA23 revealed several more regulatory sequence possibilities for both Sav1 and Sco11 and suggested that they are in monocistronic operons. An additional preceding sequence was found in common for both of these FUPA23 genes: gtcCGcacgtacCGtcg (score=6.91). Comparable regulatory sequences were also found preceding other genes in each of these, separately. A gene for an arsenical-resistance protein ACR3 (which may actually offer resistance to other heavy metals according to RegPredict) may be expressed in an operon with a gene for a putative oxidoreductase, and the pair is preceded by the sequence gtcCGcccgtacCGccg (score=5.78). A gene for a large protein (3414aa) is found to be preceded by a variant of this sequence as well in *S.avermitilis*. It is considered a putative transcriptional regulator of the SARP family (*Streptomyces* antibiotic regulatory protein), which is only known to participate in modulating the production of antibiotics in this organism [Hindra, et al., 2010]. The sequence preceding this gene is gtACGcgcgtacCGTcg. Finally, *S.avermitilis* is found to possess a gene for 4-

aminobutyraldehyde dehydrogenase (EC 1.2.1.19) (no known metal cofactors) that is preceded by the sequence gtcCGcccgtacCGccg (score=5.78).

These findings suggest that this FUPA23 is in some way connected to antibiotic production/regulation, but how is not clear. Perhaps its substrate is involved in the regulation of expression, as is seen in [Hesketh A et al. 2009]

The gene for **Efa6** (*Enterococcus faecalis* V583 806aa, gi:29375837) is found in a monocistronic operon, 147bp away from the next gene downstream and co-directional. It is however, diverging from a common candidate regulatory region of 218bp with another monocistronic operon, encoding ribonuclease H III (EC 3.1.26.4) (requires Mg^{2+} or Mn^{2+} to function, and in presence of 10 mM of Ba^{2+}, Ca^{2+}, Co^{2+}, Zn^{2+}, Cu^{2+}, Fe^{2+}, or Sr^{2+} [Ohtani N, et al., 2000]).

RegPredict analysis of the 218bp region between these genes reveals related regulatory sequences shared between *E.faecalis, E.faecium* and *L.acidophilus* preceding the FUPA23 P-type ATPase. For *E.faecalis* the exact sequence is CgTTagCttGt|tCctGaaAAtG (score=5.27) while for *E.faecium* the sequence is GagTCtctTtt|tgAgttGAaaC (score=5.15). RegPredict also suggests that the cell wall surface anchor family protein gene found 147bp downstream of this gene in the same direction may also be regulated by this sequence and thus share an operon with the gene, although the distance makes this suggestion less likely. The gene for ribonuclease HIII also comes up in RegPredict, with the regulatory sequence of CaTtgTCaTtt|tcAgGAacAaG (score=5.87). Bicistronic and monocistronic operons are

regulated by *identical* sequences (score=5.15) to that found preceding the FUPA23 gene in *E.faecalis*. The bicistronic operon encodes the genes for the two-component transcriptional response regulatory protein VncR and the sensor histidine kinase VncS which are cotranscribed and translationally coupled with a 3bp overlap. Their function is unclear, although virulence against antibiotics has been implicated (Haas W, et al., 2004). The monocistronic operon is found only in *E.faecium* and is a transposase gene of ~1138aa in size, or two separate transposases as described, one 906aa and one of 231aa with an overlap at the genetic level of 24bp. Nine such transposases are purported to be found in *E.faecium*'s genome, ranging in size from 231aa to 1296aa. Also, when these transposases are analyzed in SEED, two strains of *E.faecium* present some interesting findings. In one of them the transposase gene appears to be divergently transcribed from *vncRS.* This is the same *vncRS* pair described above. In the same strain of *E.faecium* in which a transposase gene homolog is found divergently cotranscribed with a FUPA23 gene, the two component system *phoPR* genes are found convergently transcribed from the FUPA23 homolog gene, as if the FUPA23 gene is interrupting a larger operon. The two response regulators, PhoR and VncR are homologs with 51% similarity.

The gene for **Efa10** (of *E.faecalis* V583 778aa, gi:29376085) is found as the eighth gene in a putative 13-gene operon. The genes encode the following proteins, in order of transcription:

1) muconate cycloisomerase (EC 5.5.1.1) (divalent metal ion is required, Mn^{2+} is preferred [Neidhart DJ, et al., 1990])

2) 17bp later, transglutaminase-like enzyme / putative cysteine protease
3) 18bp later, oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1)
4) 310bp later, transcriptional antiterminator bglG:CAT RNA-binding region
5) 37bp later, PTS system, N-acetylglucosamine-specific IIABC components (EC 2.7.1.69) (no known cation requirements)
6) 93bp later, 48aa hypothetical protein: FIG00632476
7) -24bp (overlap) later, pneumococcal vaccine antigen A homologue
8) 129bp later, FUPA23
9) 317bp later, DNA primase (EC 2.7.7.-)
10) 109bp later, RNA polymerase sigma factor RpoD
11) 152bp later, 1 TMS, ~690aa hypothetical protein: FIG00629252 #33% similar to Holliday junction-specific endonuclease
12) 74bp later, S1 RNA binding domain
13) 116bp later, Ferric uptake regulation protein FUR

Divergently transcribed from this operon, 276bp away, is a single gene encoding a probable aromatic ring-hydroxylating enzyme; PaaD-like protein (DUF59) involved in Fe-S cluster assembly.

RegPredict suggests that the gene for this ATPase has a regulatory sequence of AaaCgagAaAagTcTaaaGaaT (score=7.22). A similar regulatory sequence, aCacgagAaAATTaTtgaaaGa (score=5.53) may govern expression of a gene for 2',3'-cyclic-nucleotide 2'-phosphodiesterase (EC 3.1.4.16). Another gene, for Potassium

efflux system KefA protein/Small-conductance mechanosensitive channel, may be governed by the somewhat similar sequence aaacaTgttAagTcttAaaaaa (score=5.17) and, in tandem with that, aaaagaagAAATTTtagagaaa (score=4.77). Finally, a sequence of AgaCgTgAaaATgaTgAaGaaT (score=5.12) precedes a gene for rRNA small subunit methyltransferase I # 16S rRNA 2 (prime)-O-ribose C1402 methyltransferase which is followed 15bp later by a gene for YbbL ABC transporter ATP-binding protein which itself is preceded by what may be a reinforcing regulatory sequence of AaaCaagAagTAtaTaaaGaaT (score=4.84). Following the YbbL gene and overlapping it by 16bp is the gene for YbbM seven transmembrane helix protein.

The gene for **Ljo8** (*Lactobacillus johnsonii* NCC 533 809aa, gi:42519270) is preceded by a gene 17bp upstream and followed by a gene 80bp downstream, both transcribed in the same direction. The downstream gene is unlikely to be cotranscribed, but the upstream gene, which codes for a cytoplasmic L-asparaginase I (EC 3.5.1.1), most likely shares with this FUPA23 in a bicistronic operon. The asparaginase is preceded 91bp upstream by a DUF1212 integral membrane protein There is some support for use of monovalent cations by the L-asparaginase found in [Law AS et al. 1971], where it is stated that a “monovalent cation is required to effect the conversion of the enzyme-substrate complex to enzyme and product”, as well as demonstrating that maximal velocity of the enzymatic reaction is independent of the cation species. However, the concentration of cation required to effect one-half maximum velocity was shown to be species dependent in the order “potassium > lithium > sodium”. The

orthologue of this asparaginase in *L.acidophilus* is found encoded upstream of a Ca^{2+} P-type ATPase gene, separated by 3 genes for hypothetical proteins with no substantial predictive homology. It is also interesting to note that *L.gasseri* possesses a highly similar gene neighborhood for its own FUPA23 gene. In both of these organisms, FUPA23 is preceded (with ~15bp between by the cytoplasmic L-asparaginase I, but moreover, 15 neighboring genes to these two are the same. Seven of these are co-directional and upstream without antidirectional interruption, one, a PTPS related superfamily hypothetical protein, is co-directional and downstream and has 2 tandem copies in *L.johnsonii* but only 1 *L.gasseri*. This protein is predicted to have 14 TMSs and its best hit in TCDB is (TC# 2.A.39.3.1) Allantoin permease with an e-value of .018. PSI-BLAST implicates several close homologs to this protein as putative cell division proteins. The products of the 7 *co-directional* upstream genes are described in order of transcription to just before the asparaginase gene:

1) Methionine ABC transporter substrate-binding protein
2) 17bp later, Methionine ABC transporter ATP-binding protein
3) -7bp later, Methionine ABC transporter permease protein
4) 138bp later, Fumurate hydratase class II (EC 4.2.1.2)
5) 15bp later, Fumurate reductase flavoprotein subunit (EC 1.3.99.1)
6) 80bp later, L-lactate dehydrogenase (EC 1.1.1.27) (divalent cation needed, [Hensel R et al., 1977]
7) 91bp later, DUF1212 integral membrane protein; 9TMS predicted, closest TCDB hit is (TC# 2.A.79.2.1) Putative uncharacterized protein yjjP with 1e-05.

Following 91bp of intergenic space we find the asparaginase described above. Divergently transcribed from this series and 418 bp away, we find a bicistronic operon, so dubbed due to the space of 0bp between the genes involved. The genes encode the following, in order of transcription:

1) Translation initiation factor 2
2) DNA-entry nuclease (competence-specific nuclease) (EC 3.1.30.-)

Convergent with the PTPS protein gene that follows the FUPA23 and 123bp away is a gene encoding a putative transcriptional regulator YbaK/ebsC. This gene is in turn convergently transcribed from another gene 63bp away, for a Ribosomal protein L11 methyltransferase (EC 2.1.1.-). 431bp later is the gene for a GTP pyrophosphokinase (EC 2.7.6.5)/(p)ppGpp synthetase I. Immediately after this , 0bp later, is a gene for a D-tyrosyl-tRNA(Tyr) deacylase. Another 283bp later we find a Histidyl-tRNA synthetase (EC 6.1.1.21) (Mg^{2+} requirement [Airas RK et al., 1996]). Two bp later, we find an aspartyl-tRNA sythetase (EC 6.1.1.12) (Mg^{2+} requirement [Thompson D, et al., 2006]) encoded. This concludes the conserved gene cluster of FUPA23 for *L.johnsonii* and *L.gasseri*. It is also worth mentioning that the genes referred to as 1a-5a above are also found with the FUPA23s of *L.acidophilus* and *L.delbruckii* (*sans* fumurate hydratase class II), but 398bp downstream instead.

RegPredict suggests that a regulatory sequence may exist ahead of the asparaginase described above (agttgTCatgGAattta, score=6.98). Elsewhere in the genome of *L.johnsonii* is found a similar regulatory sequence (agAtgTCAaTGAatTta, score=5.42) which is suggested to regulate a 4-gene operon starting with the gene for an

ABC transporter ATPase component, followed with a 7bp overlap by a gene for a major facilitator superfamily permease, followed 50bp by the gene for the 23S rRNA (Uracil-5-) -methyltransferase RumA (EC 2.1.1.-). The last gene suggested by RegPredict to potentially be in this operon, 167bp after the methyltransferase, is a 194aa hypothetical protein with no TC hits and no informative PSIBLAST hits.

Oye3 (*Onion yellows phytoplasma* OY-M 1,056aa, gi:39938738) is described as a FUPA23 in Chan, et al. (2010), "PSI-BLAST searches revealed that the C-terminal [235aa]extension in this protein is homologous to SpoU methylases, enzymes that function in the methylation of tRNAs and rRNAs [Kuratani et al., 2008]." SEED shows this to be the case for *Aster yellows witches'-broom phytoplasma* (AYWB) and *Acholeplasma laidlawii* PG-8A d as well. However the C-terminal portion is not fused as in Oye3, and while the AYWB FUPA23 is 817 residues long, the homologous region ends at 785aa in AYWB, 788aa in OY-M. The remaining 32aa in the AYWB protein seem to be unique to this protein only. Similarly, OY-M's FUPA23 homology does not begin until the M822, at which point it is 99% identical to the SpoU methylase family protein YacO in AYWB. The 34 aminoacyl residues between these regions of homology appear unique to OY-M, though a region of 8 residues just past the center of each region is 62.5% identical. This may be sufficient to consider it an ancestral linker region from the fused ancestral protein. The unlinking of the methylase is favored over a single fusion event in OY-M because of the presence of this nonhomologous linker sequence in AYWB. While the *A.laidlawii* FUPA23 operon is not similar to Oye3's, the *Aster*

yellows operon is essentially the same. This suggests that in *O.yellows*, the FUPA23's (unfused to SpoU) SEED assignment of "Calcium-transporting ATPase (EC 3.6.3.8) / 23S rRNA (guanosine-2'-O-) -methyltransferase rlmB (EC 2.1.1.-)" is valid and may facilitate prediction of function in other organisms. The series containing Oye3 begins and proceeds as follows:

1) Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (activity depends on divalent cation, efficiency in descending order: Mg^{2+}, Mn^{2+}, Co^{2+}, $Zn^{2+,}$ $Cu^{2+,}$ Ca^{2+} [Simuth J, et al., 1975])
2) 78bp later, FUPA23 (includes C-terminal Methylase domain)
3) 223bp later, LSU ribosomal protein (RP) L33p
4) 17bp later, Preprotein translocase subunit SecE (TC 3.A.5.1.1)
5) 2bp later, Txn antitermination protein NusG
6) 198bp later, LSU RP L11p (L12e)
7) 161bp later, LSU RP L1p (L10Ae)
8) 40bp later, LSU RP L10p (P0)
9) 12bp later, LSU RP L7/L12 (P1/P2)
10) 172bp later, DNA-directed RNA polymerase β-subunit (EC 2.7.7.6) (the active center of the enzyme involves a symmetrical pair of Mg^{2+} ions that switch roles in synthesis and degradation. One ion is retained permanently and the other is recruited ad hoc for each act of catalysis. The weakly bound Mg^{2+} is stabilized in the active center in different modes depending on the type of reaction: one mode is during synthesis by the β,γ-phosphates of the incoming substrate and the other

mode is during hydrolysis of the phosphates of a non-base-paired nucleoside triphosphate [Imburgio D, et al., 2002])

11) 3bp later, DNA-directed RNA pol. β`-subunit (EC 2.7.7.6) (see above)
12) 21bp later, SSU RP S12p (S23e)
13) 89bp later, SSU RP S7p (S5e)
14) 25bp later, Translation elongation factor G
15) 131bp later, Translation elongation factor Tu
16) 572 bp later, Probable conserved transmembrane protein. This protein is not included in the operon, but is mentioned to make clear the end of the operon prior to it.

While OY-M and AYWB are rare cases of P-type ATPases found with this ribosomal large and small subunit cluster, the cluster and variants thereof are appropriately very common for the YacO rRNA/tRNA methyltransferase, being found in Bacilli, Lactobacilli, Geobacilli, Oceanobacilli, Staphylococci and Listeria. Therefore, the function of the fusion of this gene with that of FUPA23 in OY-M and AYWB as well as in *A.laidlawii* may be linked to their obligatorily parasitic lifestyle. While this unique attachment might be considered an artifact of sequencing error if t were found alone, its occurrence in three distinct but related intracellular parasites demonstrates that it is genuine. As such, and considering that this fusion is not found elsewhere, it can be assumed that the organisms benefit from this fusion related to the nature of their environment.

The gene for the protein described in Chan et al. (2010), **Stth5** (*Streptococcus thermophilus* CNRZ1066 795aas gi:55823500) is present in a gene neighborhood unique to strains of this species. Encoded preceding it is a conserved hypothetical protein with close homologues only in other Streptococci and Lactobaccili but with none having conserved domains. It is transcribed in the same direction but 233bp away from the FUPA23 ATPase gene. Downstream of the ATPase gene can be found another conserved coding region, this one coding for a protein bearing ~50% similarity to a transposase and spbAB. Very little is known about these protein products. It is transcribed in the same direction but 82bp away from the FUPA23 ATPase. Further upstream is a cluster of transcriptional regulators including OrfX, a member of the MerR family.

The gene encoding **Ter3** (*Trichodesmium erythraeum* IMS101 831aa, gi:113477109) is not found in any conserved gene cluster, being 1089bp downstream of the nearest gene in the upstream direction and 487bp separated from the next gene downstream of it.

RegPredict suggests both a major and a minor regulatory sequence precede this FUPA23 (major: gGTTattT|AtcaAACt, score=6.93; minor: tGgTatTa|aAcaAaCt, score=4.62). RegPredict suggested four similar sequences. The first precedes a gene for Molybdenum transport system permease prtein ModB (TC# 3.A.1.8.1) / Molybdenum transport ATP-binding protein ModC (gGTtgttT|AtcacACt, score=5.77). Another similar sequence (gGTTattt|gtccAACt, score=5.77) is found preceding the genes for (EC

4.1.1.21) (no cation requirements known) phosphoribosylaminoimidazole carboxylase, catalytic subunit [purE] and, 193bp later, a 12 TMS ammonium/methylammonium permease (TC# 1.A.11.2.3). The third sequence (ggTTattT|AttaAAtt, score=5.77) appears to regulate a bicistronic operon, the first gene of which is a hypothetical protein with no significant homology. Following this gene 127bp later however, a gene for a polysaccharide biosynthesis protein occurs. The fourth regulatory sequence (gATTaatTAtcaAATt, score=5.20) appears to be regulating the expression of the gene for 1,4-dihydroxy-2-naphthoate octaprenyltransferase (EC 2.5.1.-). While several of these look as if they may simply bare homology due to TATA boxes, it should be considered that the 300bp region preceding the FUPA23 has a GC content of only 23.9%.

The gene for **Tel1** (*Thermosynechococcus elongatus* BP-1 826aa, gi:22293874) is transcribed head to head with the gene for an ABC transporter ATP-binding protein. As this arrangement is not conserved in any other available organisms, no downstream-based predictions are made. Upstream of the FUPA23 however, the gene for a 195aa universal stress-type protein, UspA-like is found, separated from the FUPA23 gene by 4bp. Another 69bp before this *uspA-like* is the gene for a 240aa “putative transposase”, and 37bp before that a gene for a 141aa “ORF_ID:tll0156 putative transposase”. Divergently transcribed from this ORF_ID:tll0156, with 52bp between them, is a gene for a 116aa “ORF_ID:tlr0157 putative transposase.” While little information can be

drawn from these putative transpoases, only the largest and first mentioned possesses a zinc ribbon domain, usually implying a zinc-mediated DNA binding domain is present.

RegPredict suggests a regulatory sequence precedes the gene for the UspA-like protein (GgTgagtATgtcaAtC, score=6.93). Seven regulatory sequences similar to this were found to be predicted to regulate various operons in *T.elongatus*. The first similar sequence (GgTgagtATgtcaAac, score=6.35) appears to regulate the expression of the gene for a potassium channel of the VIC family. The second similar sequence (ggTgagtATgtcaAat, score=5.77) appears to regulate a 5-gene operon. The first gene in this predicted operon is for transaldolase (EC 2.2.1.2), followed 46bp later by the gene for a hypothetical protein of the DUF1008 superfamily which is homologous to heme utilization protein HuvX. Third in this predicted operon, 18bp later, is the gene for a cell division inhibitor, followed 12bp later by the gene for another hypothetical protein of the Fe-S_biosyn superfamily which is homologous to iron-sulfur cluster assembly accessory protein. The last gene in the predicted operon occurs 15bp later. It encodes the chorismate synthase, AroC (EC 4.2.3.5) (requires Mg^{2+} for activity according to [Hasan N, et al., 1978]). A third regulatory sequence is predicted (gGTgagtA|TgtcaACg, score=5.77) that precedes a possible operon consisting of 3 genes. The first gene encodes an exonuclease ABC subunit B, followed 158bp later by a gene encoding the phycobilisome core component protein apcF, followed finally 88bp later by a gene encoding 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (EC 4.6.1.12) (Mg^{2+} and Zn^{2+} are both required [Ramsden, et al., 2009 & Steinbacher, et al., 2002]). The fourth sequence RegPredict predicts (ggTgagtA|TgtcaAag, score=5.77) precedes an

operon fo 6 genes, the first of which encodes a 76aa hypothetical protein of the PK_C superfamily and is homologous to approximately the last fifth of a pyruvate kinase. Next, 101bp later, a 235aa hypothetical protein is encoded which is a member of the SIMPL superfamily but whose homology only allows it to be predicted to be localized to the periplasm or outer membrane. After 7bp more, we find encoded a 454aa hypothetical protein of the DUF 3370 superfamily whose only informative homologs were S-layer domain proteins. Following this 62bp later is a gene encoding the sulfur carrier protein adenylyltransferase moeB, a molybdopterin biosynthesis protein. Another 163bp after this is a gene encoding the photosystem I subunit XI, PsaL, and 51bp after this is encoded the photosystem I subunit VIII, PsaI. The fifth predicted regulatory sequence (ggTgagtATgtcaAaa, score=5.77) precedes a 6-gene operon starting with a gene encoding a 306aa hypothetical protein containing a proteasome-activating nucleotidase domain and a V-type ATPase domain and bearing homology to SMC structural maintenance of chromosomes partitioning protein. After 63bp more, the gene encoding ATP synthase gamma chain, atpC (EC 3.6.3.14) (requires Mg^{2+} [Shin, et al., 1996] is found, followed after 92bp more by UDP-glucose:tetrahydrobiopterin glucosyltransferase (EC 2.4.1.-). After 84bp, another hypothetical protein is found which encodes a 265aa protein with no putative conserved domains but with homology to SpoIID/LytB domain-containing protein. The next gene in the predicted operon has a 49bp overlap with the previously mentioned hypothetical protein gene. It encodes a putative transposase, currently just referred to as tll0382. Twelve base pairs later there is a gene encoding a SpoIID protein homolog. The sixth predicted regulatory sequence

(ggTgagtA|TgtcaAaa, score=5.77) appears to regulate a gene encoding nicotinate phosphoribosyltransferase (EC 2.4.2.11) (Mg^{2+} required [Imsande, et al., 1961]). This gene is overlapped by 3bp by a gene encoding nicotinate-nucleotide adenylyltransferase of the bacterial NadD family (EC 2.7.7.18) (Mg^{2+} required) [Imsande, et al., 1961]. This gene in turn is also overlapped by 3bp by the gene encoding nudix-related transcriptional regulator NrtR. This gene is then followed 26bp later by the gene for NAD synthetase (EC 6.3.1.5)/glutamine amidotransferase cain of NAD synthetase (requires Mg^{2+} [Ozment, et al., 1999]). Finally, the seventh predicted regulatory sequence (ggTgagtA|TgtcaAaa, score=5.77) is found preceding a predicted 7-gene operon which begins with a gene referred to as ORF_ID:tlr2239. The protein it is predicted to encode would be 115aas long but no other fruitful data could be obtained regarding its function. Following this gene 79bp later is a gene encoding DNA mismatch repair protein MutS, followed another 50bp later by a gene referred to as “ORF_ID:tlr2241 probable multidrug resistance protein”. The predicted gene product is 408aas long and is homologous to major facilitator transporters of the MFS1 superfamily (12 TMS, TC# 2.A.1.46.4). Following this putative drug efflux transporter 25bp downstream a gene encoding a “protein serin-threonin phosphatase” aka phosphoprotein phosphatase (EC 3.1.3.16) (requires Mg^{2+} for activation [Hsu J, et al., 2008]). Following this gene 16bp later is a gene only referred to as “ORF_ID:tsr2244 hypothetical protein”, which would be 59aas long but no other fruitful data could be obtained regarding its function. Following this gene 35bp later is a gene encoding cob(I) alamin adenosyltransferase (EC 2.5.1.17) (requires divalent cation, most likely Mg^{2+} [Johnson CL, et al., 2004])

Several patterns emerge upon analysis of the putative co-expression of genes with those of FUPA23s. There is a general trend towards the likelihood that they are involved in Mg^{2+} transport, or transport of a similarly applicable divalent cation. In all cases available, we see that this transport is likely in the inward direction. The SEED co-localization and RegPredict regulatory data both implicate these transporters in the import of divalent cations for the purposes translational/transcriptional enzyme metal cofactor import, possibly involving the early preparations involved in cell division. Antibiotic resistance enzyme metal cofactor import is also implicated to a lesser degree.

The FUPA24 ATPase Family

FUPA24 is a family of Type IV P-type ATPases [Chan, et al., 2010], which are all of exceptional length (1,472–1,625 aas), having a topology of 10 TMSs, with 1 & 2 pairing and found around 900 aas, 3 & 4 pairing around 1050-1150 aas and 5-10 clustered in the last 250 aas of the protein. Their unusual size is due to an ancestral fusion event of two nearly whole P-type ATPase types. The Type IV topology consists of a Type I N-terminal half and Type II C-terminal half. The family's N terminal halves' nearest hits in TCDB are Families 5 (TC# 3.A.3.5-), and 6 (3.A.3.6.-), then 25 (3.A.3.25.-) and 26 (3.A.3.26.-), depending on which one is examined. The family's C terminal halves' nearest hits in TCDB are closest to Family 2 (3.A.3.2.-). As mentioned in Chan et al. (2010), the N terminal half of these proteins is predicted to have taken on a separate role from that of an ATPase transporter, most likely macromolecular recognition. (Interestingly, the Family 2 (Ca^{2+}PA) and Family 5 (Cu^{+}PA) members related to these also independently cluster to some degree- see *Symbiobacterium thermophilum*, which has a Family 2 P-type ATPase which is the 3rd closest ATPase to the FUPA24s which is not one itself. This gene is transcribed convergently with a Family 5 gene, 2.3kb away. *B.cereus* 14579 also has a Family 2 member and a Family 6 member in close proximity). The subfamily TC# 3.A.3.24.1 is only found in Actinobacteria, all but one (*S.coelicolor*) of which are in Mycobacteriae. Three other subfamilies of FUPA24 exist that are transiently discussed in Chan, et al. (2010): Tro5 (TC# 3.A.3.24.2) is found in *Thermomicrobium roseum* (Chloroflexi), Hoc1 (TC#

3.A.3.24.3) is found in *Haliangium ochraceum* DSM 14365 (δ-proteobacteria) and Hch1 (TC# 3.A.3.24.4) is found in *Hahella chejuensis* KCTC 2396 (γ-proteobacteria). *T.roseum, H.ochraceum* and *H.chujuensis* are all currently available in SEED for analysis and thus have been included here. Orthologues of the FUPA24 genes of *M.bovis,* Mbo4 (1,539aas) and Mbo3 (1,625aas), are found in *M.tuberculosis* H37Ra (Mtu4 1,539aa, gi:148660190; Mtu3 1,632aa, gi:148659870), and *M.marinum* M (Mma4 1,537aa, gi:183980764; Mma3 1,620aa, gi:183980328) were also included in SEED analysis. *M.marinum* also possesses a third FUPA24 gene (Mma2 1,487aas, gi:183174576). As expected, the gene neighborhoods of the orthologous FUPA24 genes of *M.tuberculosis* are exactly like those of *M.bovis,* with minor exceptions. Those of *M.marinum* are highly similar as well, but with genes encoding a PAS/PAC sensor hybrid histidine kinase homologous to Stage II sporulation protein E (of *Micromonospora aurantiaca* ATCC 27029) and a putative regulatory PAS/PAC sensor protein transcribed convergently to 6a (below) of the Mbo4 ortholog and with interruptions in the neighborhood of the Mbo3 ortholog as well.

Mbo4 (*M.bovis* AF2122/97 1,539aa, gi:31791603) and its orthologs are the only FUPA24s with their genes found amongst a large number of descriptive, conserved ORFs. These conserved ORFs are also found together in *M.avium subsp. paratuberculosis str*. k10 (262316.1) without the FUPA24 gene (or any other P-type ATPase) and to a lesser degree in *M.leprae* TN (272631.1) and *N.farcinica* IFM 10152 (247156.1). Pdf, Fis, TubRel, SodC are encoded together in *M.leprae*, separately from

the FUPA24 gene, which is locally isolated. XthA, Pdf, Fis, TubRel, SodC are also found together in *N.farcinica*, also apart from the FUPA24 gene. Starting the general cluster from the most upstream gene co-directional to the FUPA24 gene, the gene cluster encodes as follows:

1) Peptide deformylase (EC 3.5.1.88), Pdf (in order of cofactor activity: Mn^{2+}>Fe^{2+}>Cu^{2+}>Zn^{2+} [Beyer,W et al., 1991]) all of these are transcribed co-directionally to upstream of FUPA24's operon, going from nearest to most distant
2) 0bp later, a gene encoding a protein homologous to GCN5-like N-acetyltransferase as well as Histone acetyltransferase HPA2 and related acetyltransferases (EC 2.3.1.48) (Ca^{2+}, Mg^{2+} [Wiktorowicz JE, et al., 1982])
3) 3bp later, Exodeoxyribonuclease III, XthA, (EC 3.1.11.2) (Mg^{2+} is the predicted metal cofactor, Mn^{2+} and Co^{2+} can substitute, see FUPA23 Blo5, 5a)
4) 201bp later, a 2 TMS hypothetical protein, with no significant homology
5) 52bp later, FUPA24 (not found in this cluster in *M.avium, M.vanbaaleni, M.smegmatis*)
6) 50bp later, hypothetical protein Rv0424c with no significant homology
7) 152bp later, Thiamin biosynthesis protein, ThiC
8) 27bp later, Phosphmethylpyrimidine Kinase, ThiD (EC 2.7.4.7) (*might* bind Mg^{2+} as seen in *S.cerevisiae* [Lewin LM, et al., 1961])
9) -3bp (overlap) Esterase/lipase/thioesterase family active site

10) 161bp later, Mycobacterium-specific FUPA24-localized possible transmembrane (TCDB predicts 2 TMSs) 137aa protein

Transcribed divergently, 133bp away from the peptide deformylase gene, the following proteins are encoded, in order of transcription:

1) putative Fis family transcriptional regulator (Fis)
2) Tuberculin related peptide (name given to extracts of *M.tuberculosis*, *M.bovis*, or *M.avium* that is used in skin testing in animals and humans to identify a tuberculosis infection, TubRel)
3) Periplasmic superoxide dismutase [Cu-Zn] precursor SodC (EC 1.15.1.1) (*may* bind $1Cu^{+}$ and $1Zn^{+}$ each [Beyer W, et al., 1991])
4) carboxylate-amine ligase (missing in the *M.leprae* cluster only)
5) Uncharacterized protein, similar to the N-terminal domain of Lon protease

RegPredict analysis of the ~133bp region between these divergently transcribed clusters in Mycobacteria suggests double regulatory sequences preceding the peptide deformylase gene upstream of the FUPA24 gene in *M.bovis* and *M.tuberculosis*, and a single such sequence in *M.marinum*. The sequences are predicted in pairs, which almost entirely overlap. They are (CGCcgCcC|c|GcGgcGCG, score=6.92) 90bp upstream and (CGCcgCctcgcGgcGCG, score=6.40) 49bp upstream as well as the overlapping sequences (CcgCcgccccgcGgcG, score=6.78) 92bp upstream and (CcgCcgcctcgcGgcG, score=6.25) 47bp upstream. The related sequence ccgCcgcccagcGgca (score=5.99) is found 53bp upstream of a gene encoding Manganese transport protein MntH which is followed 1bp later by a gene encoding a 354aa hypothetical protein with 84% positive

homology to phage-related replication protein [*M.avium subsp. paratuberculosis* S397], length=359. Finally, another related sequence, acgCcgcccagcGgcg (score=5.99) is found 88bp upstream of a gene for a probably exported protease (EC 3.4.-.-) which is followed 95bp later by a gene for a glutamine synthetase type I (EC 6.3.1.2) (requires two divalent cations, either Mg^{2+} or Mn^{2+}, possibly depending on isomer, pH variation or transferase/biosynthetic activity [Gill HS, et al., 2002; Wedler FC, et al., 1980]

The other FUPA24 in *M.bovis*, **Mbo3** (*M.bovis* AF2122/97 1,625aa 31791285), also has orthologs in *M.tuberculosis* and *M.marinum*, as well as in the single FUPA24s found in *M.avium* and *M.leprae*. Mbo3 and Mtu3 will be described together. No co-directional genes are found immediately downstream of these FUPA24 genes, as they are transcribed, head-to-head (95bp apart) from a gene for a putative GTPase of the G3E family, possessing the C terminal domain motif of CobW and thus implicated as zinc limitation response/zinc regulated enzymes. Farther downstream, approximately 110bp, an LSU ribosomal protein L28p is sometimes encoded. Strains of both *M.bovis* & *M.tuberculosis* are sometimes lacking this gene in thus cluster. In either case, either head-to-head with the L28p gene if present, 149bp away, or 543bp if not, a cAMP-dependent protein kinase is found. Tail-to-tail with this gene, 144bp away, we find the Family 5 Cu^{+}-transporting P-type ATPase Mbo2 (*M.bovis* AF2122/97 752aas gi:31791281) encoded. In *M.bovis, tuberculosis and marinum* can be found upstream of and co-directional with the FUPA24 gene, 354bp away, a hypothetical protein Rv0108c, which lacks significant homology with any non-Mycobacterial proteins (also present

adjacent to *M.avium*'s FUPA24). Further upstream still, and divergently transcribed from the Rv0108c gene, is encoded a PE-PGRS family protein (*M.avium* lacks further homology here), then 148bp later, a gene encoding rhomboid family integral MB protein/putative Ser protease. The next gene downstream encodes a transmembrane acyltransferase (EC 2.3.1.-). This gene and those immediately downstream of it are not present in the FUPA24 gene cluster of *M.marinum*. This acyltransferase is also the next gene upstream from the FUPA24 gene of *M.leprae*, 4.7kb away.

RegPredict analysis using the first 300bp of intergenic space upstream of the FUPA24 gene in *M.bovis*, *M.tuberculosis* and *M.marinum* suggests that the palindromic sequence CgCcctAtCggtaGtTgaaGgG (score=7.72) found 29bp upstream of the FUPA24 genes in *M.bovis* and *M.tuberculosis*, and the related sequences CgCggtAcacgttatTgaaGgG (score=7.02) found 28bp upstream of fupa24 in *M.marinum*, CgCggtAgcgGCgatTgaaGgG (score=6.18) found 28bp upstream of fupa24 in *M.leprae* and CCgcggAgacgtaatTgaagGG (score=4.83) found 85bp upstream in *M.avium* regulate expression of their respective FUPA24 genes. Related sequences in *M.marinum* (CgCcgtGttGgtCggCcaaGgG, score=4.80), as well as *M.bovis* and *M.tuberculosis* (cgCggtGttggtaggCgaaGgt, score=5.25) are found preceding a conserved operon consisting of several of the genes involved in various biosynthetic processes. The genes that appear to be under the control of the regulatory sequence (in downstream order) encode Tryptophan synthase alpha and beta chains, TrpA and TrpB (both EC 4.2.1.20, genes have 3bp overlap) followed in *M.bovis* and *M.tuberculosis* 0bp later by *lgt* which encodes prolipoprotein diacylglyceryl transferase (468aa) and in *M.marinum* by a gene

for the same protein, but with an overlap of 3bp and a length of 714aa, these proteins have a low specificity monovalent cation requirement [Dierkers AT, et al., 2009]). In *M.bovis* and *M.tuberculosis*, this is the end of the operon. Following this gene in *M.marinum* however, we find a gene encoding a hypothetical protein 64bp downstream, which is only found to be homologous to a TM2 domain containing protein, which is of unknown function. Following this gene with a 13bp overlap is encoded another hypothetical protein, of which there is no significant homology and little is known except that it is a probable membrane protein with 3 TMSs. Prior to the genes thus far mentioned in *M.marinum*, we find genes encoding anthranilate synthase component I TrpAa (EC 4.1.3.27) (requires Mg^{2+} [Zalkin H, et al., 1968]) followed 41 bp later by a gene encoding a 4 TMS hypothetical protein homologous to "trp region conserved hypothetical membrane protein". Another 90bp after this, we find encoded Indole-3-glycerol phosphate synthase TrpC (EC 4.1.1.48) (Ca^{2+}, Mg^{2+}, Mn^{2+} and Na^{+} "significantly affects activity, the optimal concentration is about 0.4-2.0 mM" according to Yang Y, et al. (2006), in which is found the regulatory sequnce suggested above, 112bp from its end, which is itself 36bp upstream of *trpB*. While the regulatory sequence suggested above is not found prior to these genes in *M.avium* or *M.leprae*, it is suggested that another regulatory sequence may also effect expression of this operon, since the operon is found in these organisms. In fact the entire operon descibed for *M.marinum* is found as well in *M.avium* and *M.leprae*, except that *M.leprae* lacks the second hypothetical protein and *M.avium* possesses genes for a pyruvate kinase (EC 2.7.1.40) 130bp downstream of it, followed 73bp later by a gene for Acyl-CoA thioesterase II (EC

3.1.2.-) and then another 142bp later, a putative membrane protein with 38% homology to integral membrane lysyl-tRNA synthetase.

Mav7 *(M.avium subsp. paratuberculosis* K-10 1611aa, gi:41409596) is not co-localized with anything except the previously mentioned Mycobacteria-specific Rv0108c protein.

RegPredict predicts 2 other regulatory sequences shared with **Mle4** (*M.leprae* TN 1609aa, gi:15828441) (tttaggCgcgGtagcgg, score=6.65): one regulating DUF1794 (tttaTACgctGTAccgg, score=6.15) and one regulating the operon starting and proceeding as follows:

1) *Pantanoate--beta-alanine ligase (EC 6.3.2.1) ($Mg^{2+/}Mn^{2+}>Ni^{2+}>Co^{2+}$ [Zheng R, et al., 2001])

2) *Aspartate 1-decarboxylase (EC 4.1.1.11) (No Metal Cofactors)

3) *Pantothenate kinase type III, CoaX-like (EC 2.7.1.33) (Cofactor flexible; Hong BS, et al. (2006) demonstrated activation by K^+, Rb^+ and most highly of the monovalent cations, NH_4^+, and also found a requirement for Mg^{2+}, which can be replaced by Co^{2+} or Mn^{2+} (possibly with stronger activation than with the native Mg^{2+}) or Zn^{2+} with 50% of the strength of activation with Mg^{2+})

4) Lysyl-tRNA synthetase (class II) (EC 6.1.1.6) (requirement for $2Mg^{2+}$, and possibly also for NH_4^+ or other monovalent cation according to Hele P, et al., 1972)

5) ”Histone protein Lsr2” (has a second, weak reg seq, max score 5.23, sequence tttggAcgcgtTagcgg).

*The first 3 of these are involved in Coenzyme A biosynthesis.

As mentioned above, *M.marinum* also possesses a third FUPA24 gene, **Mma2** (1487aas, gi:183174576). This gene appears to have one other gene that is locally codirectionally transcribed, found 133bp downstream, which encodes 212aa hypothetical protein with almost no significant homology. The only informative hit found in PSIBLAST was AzicU6 [Kibdelosporangium sp. MJ126-NF4], a 139aa protein with which it has 50% positive hits which is predicted to be involved in the biosynthesis of azicemicins [Ogasawara Y, et al., 2010] based on localization of its coding sequence in the Azic gene cluster, although no specific function is predicted for AzicU6. Divergently transcribed from this FUPA24 gene, we find, 524bp away, four genes in series with much more information. First is a gene encoding a phosphate transport system regulatory protein, PhoU, 72bp later we find a gene encoding a Magnesium and Cobalt transport protein, CorA. After 205bp more a gene encoding a ribosomal large subunit pseudouridine synthase A (EC 4.2.1.70) (dependent on presence of Mg^{2+}, Co^{2+}, Fe^{2+} or Mn^{2+}, inhibited by Zn^{2+} and Ni^{2+} [Heinrikson RL, et al., 1964; Preumont A, et al., 2008]), finally followed 22bp later by a gene encoding an Arylsulfatase (EC 3.1.6.1) (requires Ca^{2+}, which can be partially substituted instead with Ba^{2+}, Mg^{2+}, or Sr^{2+} [Moriya T, et al., 1980]).

RegPredict suggests several regulatory sequences controlling this protein that can be found elsewhere in the genome. The sequence gcgGCGcTAaCGCatt (score=6.93) is found in front of the FUPA24 itself, 95bp upstream. The related sequence

gcgGCGCaaGCGCatg (score=5.20) is found 160bp upstream of a gene encoding UDP-glucose 4-epimerase GalE2, which is followed 6bp later by a gene encoding a conserved acetyltransferase protein. Similar sequences appear to regulate this bicistron in *M.bovis*, *M.vanbaalenii* and Rhodococcus. Another similar sequence, gtgGCGCt|cGCGCact (score=5.20) is found 64bp upstream of a gene encoding isoleucyl-tRNA synthetase IleS. Lastly for this sequence group, the sequence gtgGcGct|taCcCgtt (score=5.20) is found 138bp upstream of a gene encoding a PE-PGRS family protein. Another interesting predicted regulatory sequence occurs 71bp upstream of the FUPA24 gene; it is gtGCgaCCggGGagGCtg (score=6.63). The comparable sequences gcGCgaaCggGcagGCtg (score=5.31) and gtGCgatcgacgcgGCcg (score=5.89) are found 101bp and 40bp upstream, respectively, of a bicistronic operon consisting of a gene encoding a 296aa hypothetical protein which has its closest homologue in a heavy metal-associated domain protein 313aa [Aeromicrobium marinum DSM 15272] with 58% positives, although the hypothetical protein itself does not appear to possess any conserved domains and has only a single predicted TMS. Perhaps more interesting is the protein-encoded 11bp downstream of this: a 790aa, 8 TMS P-type ATPase Cu^{+} Transporter (TC# 3.A.3.5.18). A similar regulatory sequence, gcGCggtcg|acgaaGCcg (score=5.03) is found 40bp upstream of an orthologue in *M.avium* as well. This regulatory sequence pattern also predicts a second operon, which may be regulated by two adjacent sequences. The sequences are gtgggTtcggtgAgggtg (score=4.68) and gtGcgaCcggtGcgaCcg (score=5.51) and they occur 192bp and 150bp, respectively, of their associated genes. The genes they are predicted to regulate are predicted to be in a 5-

gene operon according to RegPredict. The first gene encodes a 207aa hypothetical protein with a 93% positive score with its homologue, DsbA oxidoreductase [*M.parascrofulaceum* ATCC BAA-614], length=207aa. Following this gene 132bp downstream is a gene encoding C4-dicarboxylate-transport transmembrane protein DctA, and after another 86bp we find encoded a Ribose-5-phosphate isomerase B (EC 5.3.1.6) (no activation by mono- or divalent cations, [Park CS, et al., 2007] RpiB/Galactose 6-phosphate isomerase, followed 9bp later by a gene encoding formamidopyrimidine-DNA glycosylase (EC 3.2.2.23) (requires Zn^{2+} bound in its zinc-finger motif for proper damaged DNA recognition, Co^{2+} can also bind, [Buchko GW, et al., 2000] which is finally followed 128bp later by a gene for a 119aa hypothetical protein which is proline-rich throughout and very highly glycine-rich in the last 3rd of the protein and is predicted by TCDB to have a single TMS near the N-terminus. The next predicted regulatory sequence is gcGcgaCcggcGcgcCtg (score=5.51) and is 94bp upstream of the genes it is predicted to regulate: first a gene encoding a membrane-anchored adenylyl cyclase, followed 101bp by a gene encoding a 127aa hypothetical protein which is a 92% positive match with pyridoxamine 5-phosphate oxidase [*M.parascrofulaceum* ATCC BAA-614], length=128aa. Finally, the last predicted regulon for this sequence pattern to be discussed is preceded by the sequence gTGCcaCcggtGatGCAg (score=5.19) and it is found 59bp upstream of the first gene in the 4-gene cistron it is predicted to regulate. This gene encodes a 318aa hypothetical protein which, interestingly, is a 75% positive match with putative von Willebrand factor, type A [Saccharopolyspora erythraea NRRL 2338], length=315. This gene is

followed 35bp later by one encoding a 335aa membrane protein with 5 TMSs also has its first relevant PSIBLAST hit in von Willebrand factor [*M.parascrofulaceum* ATCC BAA-614], length=335, 96% positive. However this protein also has a second informative hit in Mg-chelatase subunit ChlD [*M. sp.* Spyr1], length=335, 92% positive. This gene is followed 114bp later by a gene encoding 3-oxoacyl-[acyl-carrier protein] reductase, FabG1 (EC 1.1.1.100) (binds NADP+, does not use metal cofactor [Silva RG, et al., 2006 & 2008]), which is then followed by a gene encoding enoyl-[acyl-carrier protein] reductase, InhA (EC 1.3.1.9) (activation by univalent cations in descending order, NH_4^+, Rb^+, Cs^+, K^+, and Na^+ [Marrakchi H, et al., 2003]).

Nfa7 (*Nocardia farcinica* IFM 10152 1598aa* gi:54024683) is encoded with intergenic spaces upstream and downstream 340bp and 640bp long, respectively. It is therefore predicted to be a monocistron, although it may be expressed with an operon upstream of it which is divergently transcribed, as well as being coregulated with genes elsewhere in the genome. The downstream neighborhood of this gene is comprised of hypothetical proteins with no significant homology to known proteins or domains (the closest match is 35% similarity to a Methyl-accepting chemotaxis protein). The upstream neighborhood of this gene however is more coherent. Divergently transcribed from the FUPA24 gene is encoded a two component transcriptional regulator of the winged helix family. Downstream of this gene by 198bp is an operon of ectoine biosynthesis and regulation proteins starting with a gene encoding L-2,4-diaminobutyric acid acetyltransferase (EC 2.3.1.-), EctA. Overlapping this gene by 3bp is a gene encoding

diaminobutyrate-pyruvate aminotransferase (EC 2.6.1.46) (most likely requires no heavy metal ions [Rao DR, et al., 1969]), EctB which is in turn overlapped by 3bp by a gene encoding L-ectoine synthase (EC 4.2.1.-), EctC which is followed 16bp later by a gene encoding ectoine hydoxylase (EC 1.17.-.-), EctD.

RegPredict suggests a regulatory sequence preceding this FUPA24 gene by 186bp of aCGgcGcgcgcCcgCGc (score=5.89). A similar sequence, tCGgcGcaggcCcgCGc (score=4.54) is found 153bp upstream of a gene encoding aspartate aminotransferase (EC 2.6.1.1) (no metal cofactors known). Interestingly, elsewhere in the genome similar genes, both encoding L-lactate dehydrogenase [cytochrome] 1 (EC 1.1.2.3) (this protein is a ferricytochrome and thus binds Fe^{2+} in the form of a heme cofactor [Lê KH, et al., 2009]) are found. The first is included in an operon preceded 94bp upstream by the sequence GcGtGGGCgGCCCgCcC (score=4.60). The first gene in this bicistron encodes a 233aa hypothetical protein in the creatininase superfamily, its closest hit being Creatininase [Frankia symbiont of *Datisca glomerata*], length=240 with a 67% positive score. This gene is overlapped 3bp by the gene encoding one of the aforementioned L-lactate dehydrogenases. The other is the first in its operon, being preceded 118bp by the sequence ccGgcGcgcgcCcgCcc (score=5.39). This L-lactate dehydrogenase gene is itself overlapped 3bp by a gene encoding a glycosyl transferase. Still another related sequence, GCGgCGcCcGcCGaCGC (score=5.22) is found 238bp upstream of a gene encoding a 494aa hypothetical protein with closest known hits of GTPase of unknown function [*Saccharomonospora viridis* DSM 43017], length=617 (45% positive score) and HSR1-like GTP-binding protein

[*Actinosynnema mirum* DSM 43827], length=610 (45% positive score), though it lacks any conserved domains. The related sequence acGtGGcgggcCCgCcc (score=5.20) precedes a bicistron by 31bp. The first gene encodes a low molecular weight protein tyrosine phosphatase ptpA (EC 3.1.3.48) (Cu^{2+}, Zn^{2+} and Mn^{2+} may activate, Mg^{2+} and Ca^{2+} may have no effect [Mijakovic I, et al., 2005] followed 0bp later by a gene encoding cytochrome oxidase biogenesis protein Surf1 (which facilitates heme A insertion). Finally, a sequence of acGgcGcgggcCcgCcg (score=5.17) is found 119bp upstream of a gene encoding argininosuccinate synthase (EC 6.3.4.5) (no known metal cofactor requirements), which is then followed 33bp later by a gene encoding argininosuccinate lyase (EC 4.3.2.1) (no known metal cofactor requirements).

Sco5 (*S.coelicolor* A3(2) 1473aa* gi:21220960) (*is misrepresented as having only 760 aas in Chan, et al. 2010, NCBI and SEED both show this protein to be 1472aa long.) Depending on which strain of *S.coelicolor* is examined, the gene for this FUPA24 is either 26bp apart divergently transcribed from a putative membrane protein with no significant homology which is only found in Streptomycetes next to this FUPA24 gene, or the FUPA24 gene is the third gene in an operon consisting of (in order of transcription) a gene for a putative DNA-binding protein, followed 182bp later by a gene encoding a putative histone-like DNA-binding protein, which is then followed 10bp later by the FUPA24 gene. The putative DNA-binding protein gene and all else following are common to both strains, including their location relative to the FUPA24 gene. Transcribed divergently from the putative DNA-binding protein gene, 111bp away, is a

gene encoding a putative transmembrane efflux protein containing a MFS Family domain and very strong homology to an actinorhodin transporter. Downstream of this gene by 123bp is one encoding a putative endochitinase precursor. On the other side of the FUPA24 gene, convergently transcribed with 40bp away, is a homolog of the putative transmembrane efflux protein containing a MFS Family domain just mentioned, this one however is strongly homologous to the EmrB/QacA family drug resistance transporter.

RegPredict suggests a regulatory sequence of acGGcCaCcGgGtCCac (score=6.34) preceding the FUPA24 gene. Similar sequences precede and are predicted to regulate several other genes, including an operon methionine transport and possibly metabolism. The sequence is GCGGcCaCcGcGtCCGC (score=5.17) and the proteins encoded are descibed as follows in order of transcription.

1) Methionine ABC transporter permease protein
2) Methionine ABC transporter substrate-binding protein
3) putative acetyltransferase
4) Cobalt-precorrin-6y C5-methyltransferase (EC 2.1.1.-)

Another regulatory sequence (aCGGcgaCcGcgaCCGc, score=5.17) precedes a gene encoding an ATP-dependent Clp protease proteolytic subunit (EC 3.4.21.92) (“ClpA degrades large proteins down to short peptides of 7 to 10 amino acids, the process requires both Mg^{2+} and ATP hydrolysis” Mn^{2+} can substitute [Porankiewicz J, et al., 1999]).

The regulatory sequence ccGGcCgCcGgGtCCac (score=5.22) precedes a lone gene encoding the protein L-fuconolactone hydrolase.

A regulatory sequence of tcGGcCCtcgGGaCCac (score=5.02) precedes a 4-gene operon encoding first a thiamine pyrophosphate-requiring acetolactate synthase (EC 4.1.3.18) (no known ion requirements), then a putative formyl-coenzyme A transferase, then a NDP forming Acyl-CoA synthetase and finally a transmembrane transport protein of the MFS superfamily which is homologous to Oxalate/Formate Antiporter, with a 49% positive score and e-value of 4e-47.

Another similar sequence (acGacCcCcGtGtcCac, score=5.10) precedes another 4-gene operon encoding first (#1) 6,7-dimethyl-8-ribityllumazine synthase (riboflavin synthase beta subunit), ribH (EC 2.5.1.9) (no ion requirement demonstrated in Bacteria), then (#2) phosphoribosyl-ATP pyrophosphatase, hisE (EC 3.6.1.31) (requires Mg^{2+} [Javid-Majd F, et al., 2008]), then (#3) ATP phosphoribosyltransferase, hisG (EC 2.4.2.17) (requires Mg^{2+} [Lohkamp B, et al., 2004]) and finally (#4) a putative membrane protein with no significant homology.

Another related sequence (GcGacCaCcGcGtgCaC, score=5.17) is predicted to regulate a 5-gene operon encoding a (#1) 250aa hypothetical YebC-like protein, then (#2) crossover junction endodeoxyribonuclease, RuvC (EC 3.1.22.4) (requires Mg^{2+} or Mn^{2+} for cleavage mechanism [Takahagi M, et al., 1994]), then (#3 & 4) Holliday junction DNA helicases, RuvA and RuvB and lastly (#5) preprotein translocase YajC (TC# 3.A.5.1.1).

The related sequence (aCGGcCaCcGcGaCCGg, score=5.17) is predicted to regulate another 5-gene operon encoding first (#1) D-amino-acid oxidase (EC 1.4.3.3) (no known metal ion requirements in prokaryotes), then (#2) a hypothetical protein similar to sarcosine oxidase alpha subunit, possessing a 2Fe-2S domain, then (#3) putative FAD-dependent pyridine nucleotide-disulfide oxidoreductase, then (#4) 1-pyrroline-4-hydroxy-2-carboxylate deaminase aka dihydropicolinate synthase (EC 3.5.4.22) (no known metal ion requirements) and finally (#5) 4-hydroxyproline epimerase (EC 5.1.1.8) (no known metal ion requirements).

Finally, the related sequence, acGGcCaccccGaCCac (score=5.75), is predicted to regulated a bicistron which encodes first a 121aa hypothetical protein with homology to putative conjugal transfer protein [*S.himastatinicus* ATCC 53653] Length=117, with E-value=2e-55 and Positives=91% and then second (#2) a NADP-dependent, aldo/keto reductase family oxidoreductase.

Tro5 (TC#3.A.3.24.2 *T.roseum* DSM 5159 (Chloroflexi) 1607aa, gi:221635885), **Hoc1** (TC#3.A.3.24.3 *H.ochraceum* DSM 14365 (δ-proteobacteria) 1441aa, gi:262193822) and **Hch1** (TC#3.A.3.24.4 *H.chejuensis* KCTC 2396 (γ-proteobacteria) 1446aa, gi:83643329) are interesting cases because even as representative members of their respective subfamilies of FUPA24 proteins, they are the only ones found in their those phyla/classes rather than Actinomycetes. Other organisms possessing homologous proteins which are hits as FUPA24s in TCDB include *Thermobispora bispora, Streptosporangium roseum* and *S.arenicola* and *Frankia sp.* Ccl3 and EAN1pec.

In *T.roseum*, the FUPA24 gene is found in a large gene neighborhood of relatively small hypothetical proteins with a few others with defined functions. Three of the hypothetical protein genes are predicted to be in an operon with the FUPA24 gene, but 4 others are also found nearby. All hypothetical proteins were run through NCBI- and TC-BLAST to check for significant homology, none was found. Two of these proteins are found immediately upstream of the FUPA24 gene, transcribed co-directionally, one 22bp away from it and the other 14bp apart from that. Another is transcribed divergently from this, 61bp away. Convergently transcribed with 5bp overlapping from this, but in the same direction as the FUPA24 gene, is encoded Mannose-6-phosphate isomerase (EC 5.3.1.8) (require a divalent ion metal cofactor for activity and catalysis [Yeom SJ, et al., 2009]). This gene is most likely the 4th and last in an operon which begins encoding Dihydrolipoamide dehydrogenase of branched-chain alpha-keto acid dehydrogenase (EC 1.8.1.4) (no metal ion requirements known in prokaryotes), then Na^+/H^+ antiporter (TC# 2.A.37.2.4), then a hypothetical protein. One hypothetical protein gene 141bp downstream of the FUPA24 gene is also predicted to be transcribed with it. Another hypothetical protein gene is found 345bp downstream of this, transcribed codirectionally. Convergently transcribed 65bp away from this is a gene encoding a Fumarylacetoacetate hydrolase family protein. A gap upstream of this gene of 1170bp precludes any further coexpression.

RegPredict suggests a regulatory sequence of cGtgGcaGCgCGgGCacCagCc (score=6.96) upstream of the FUPA24 gene. A similar sequence of

aGggGaaGcaCGatCacCtgCc (score=5.03) is predicted to regulate a 12-gene operon encoding as follows (in order of transcription):

1) ATP synthase F1, epsilon subunit, atpC
2) 12bp later, UDP-N-acetylglucosamine 1-carboxyvinyltransferase, murA
3) 32bp later, c-type cytochrome biogenesis protein/thioredoxin, dsbE
4) 13bp later, 159aa hypothetical protein with 61% positive homology to copper resistance D [*Sphaerobacter thermophilus* DSM 20745], length=160
5) 33bp later, cytochrome c-type biogenesis protein, CycJ
6) 17bp later, cytochrome c-type biogenesis protein, CycK/CcmF
7) 44bp later, cytochrome c-type biogenesis protein, CcmH
8) 0bp later, 126aa hypothetical protein with no significant homology
9) -3bp later (overlap), heme ABC exporter, ATP-binding protein CcmA
10) -3bp later (overlap), heme exporter protein B
11) 35bp later, heme exporter protein C
12) 25bp later, 63aa hypothetical protein with no significant homology

Another related sequence (aGtggcTGCtCGgGCAcaagCg, score=5.03) is predicted to regulate a 3-gene operon encoding (#1) glycosyl transferase, group 1 family protein, then (#2) 130aa hypothetical protein with no significant homology and then (#3) 433aa hypothetical protein with no significant homology.

Three lone genes are also found to be preceded by related predicted regulatory sequences. The sequence CttgGcagCaCGgGtacCagcG (score=5.15) is predicted to regulate a gene encoding fructose-1,6-bisphosphate, class II. The sequence

CGggGaaGcgTAgcCagCagCG (score=5.03) is predicted to regulate a gene encoding phospho-2-dehydro-3-deoxyheptonate aldolase. Lastly, the sequence cgCggaaGCgctgGCaggaGat (score=5.03) is predicted to regulate a gene encoding a 248aa hypothetical protein with strong homology (e-value=1e-34) to putative transmembrane anti-sigma factor [*S.thermophilus* DSM 20745], length=265.

In *H.ochraceum*, the FUPA24 gene is found with both 2 downstream and 2 upstream local co-directional genes. If this cluster is treated as an operon, the order of transcription, by product name, is as follows:

1) serine/threonine kinase PKN9
2) 138bp later, a 339aa hypothetical protein with one TMS is encoded. Its best informative hit in NCBI PSIBLAST is PEGA domain protein [*Anaeromyxobacter dehalogenans* 2CP-1] Length=338, Score=162 bits (409), Expect=2e-44
3) 129bp later, FUPA24 is encoded
4) 258bp later, TPR repeat containing exported protein; putative periplasmic protein contains a prenylyltransferase domain
5) 12bp later, peptidoglycan-binding LysM

Two genes are also transcribed divergently to that of serine/threonine kinase PKN9, 129bp away. The first encodes a 142aa glyoxalase/bleomycin resistance protein/dioxygenase, which appears to have one TMS. This gene is followed by the next with a 3bp overlap, confirming co-expression. The second gene is 169aa SNF2

superfamily protein. While neither of these proteins are predicted to be of conserved co-localization with the FUPA24 gene, they are both found often with the divergently transcribed serine/threonine kinase PKN9 as well as another Ser/Thr kinase, PKN3 (EC 2.7.11.1) (no activity with cations other than Mg^{2+} or Mn^{2+} [Sharma K, et al., 2004], no activity with Mg^{2+} alone, no activity with Cu^{2+} or Zn^{2+} [Gopalaswamy R, et al., 2004]).

Using the 129bp intergenic region between the divergent genes mentioned above as a query, RegPredict suggests a regulatory sequence of gCGcCcGgcaCcGcCGg (score=6.29) preceding the FUPA24 gene. An operon consisting of 5 genes is predicted to be regulated by the similar sequence gCGcCgcgcacgGcCGa (score=6.01). These genes, in order of transcription, encode the following products:

1) GCN5-related N-acetyltransferase
2) pyridine nucleotide-disulphide oxidoreductase dimerization region
3) hypothetical protein with no significant homology
4) Na^{+}/solute symporter
5) Tetratricopeptide TPR 2 repeat protein; another tetratricopeptide, TPR 4 is also predicted to be regulated by a similar sequence, gCGccGcGcCcCacCGg (score=5.56).

Several similar potential regulatory sequences appear to regulate serine/threonine protein kinases. The sequence gCGcCGCGcCGCGcCGa (score=5.88) is found preceding a single gene which encodes one, another sequence (gCGctGCgcaGCgcCGg, score=5.56) precedes a gene encoding one which is followed 182bp later by a gene for a transketolase, and a pair of overlapping sequences (GCGcCGCGcCGCGcCGC,

score=5.51 and gCgcCGCGcCGCGcgGa, score=5.29) precede the following 4-gene operon which includes an exceptionally large one (1414aa). The operon in order of transcription consists of genes encoding the following:

1) porin LamB type Maltoporin-like (maltose/maltodextrin high-affinity receptor, phage lambda receptor-like protein)
2) RNA polymerase sigma-24 subunit, ECF subfamily
3) serine/threonine protein kinase
4) hypothetical protein with no significant homology

The sequence gCGcCcCgcgGcGcCGg (score=5.78) is suggested to regulate a bicistron encoding a major facilitator superfamily protein MFS 1 and (127bp later) an UDP-glucuronosyl/ UDP-glucosyltransferase. The sequence gCGcCgcgcgcgGcCGg (score=5.65) is predicted to regulate a monocistron encoding a protein in the benzoate-CoA ligase family. The sequence gCGcCGtgcacCGcCGa (score=5.65) is predicted to regulate a bicistron encoding PDZ/DHR/GLGF domain protein and type IV pilus assembly PilZ. Another similar sequence (gCGcCCagcagGGcCGg, score=5.65) is suggest to regulate a 6-gene operon encoding (in order of transcription):

1) CDP-alcohol phosphatidyltransferase
2) hypothetical protein, no significant homology
3) hypothetical protein, no significant homology
4) response regulator receiver protein
5) pseudouridine synthase, RluA family
6) hypothetical protein, no significant homology

The related sequence gCGcCcggcatcGcCGa (score=5.65) is predicted to regulate the 4-gene operon encoding (in order of transcription)

1) biopolymer transport protein ExbD/TolR
2) hypothetical protein, no significant homology
3) WD40 domain protein beta Propeller
4) Ppx/GppA phosphatase (EC 3.6.1.11) (divalent cation required, Mg^{2+} is most effective [Bolesch DG, et al., 2000])

A similar sequence (gCGcCcGGcCCcGtCGg, score=5.56) is predicted to regulate a 4-gene operon encoding (in order of transcription) the following:

1) Hypothetical protein with 49% positive homology to polyhydroxyalkanoic acid system protein [*Stenotrophomonas maltophilia* R551-3]
2) TM2 domain containing protein
3) RNA polymerase, sigma-24 subunit, ECF subfamily
4) hypothetical protein, no significant homology

A pair of related but non-overlapping sequences are predicted to regulate a gene encoding rRNA (guanine-N(1)-)-methyltransferase. The sequences are gCGGCCcacgcGGCCGg (score=5.06, 68bp upstream of the gene) and gCGGCGCgcgGCGCCGg (score=5.51, 32bp upstream of the gene).

Lastly, the sequence gCGGCgcgctcgGCCGg (score=5.51) is predicted to regulate a gene encoding a metallophosphoesterase.

Hch1 (TC#3.A.3.24.4 *H.chejuensis* KCTC 2396 (γ-proteobacteria) 1446aa, gi:83643329) is encoded by a gene which has no local co-directional genes downstream but 3 that are upstream, the first 2 of which are involved in the Entner-Doudoroff Pathway. If we treat this cluster as an operon, transcription would begin and proceed as follows:

1) NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) (no specific ion needs described)
2) 26bp later, Pyruvate kinase (EC 2.7.1.40) (Mg^{2+} kinetics requirement; Mn^{2+}can partially replace Mg^{2+}, other divalent cations are inhibitory; in decreasing order of inhibitory efficiency they are: Ni^{2+}, Zn^{2+}, Cu^{2+}, Ca^{2+}, Ba^{2+} [Kapoor R, et al., 1983]).
3) 305bp later, a 232aa COG1801: uncharacterized conserved protein that appears to be homologous to a sensor histidine kinase
4) 97bp later, the FUPA24 gene

Divergently transcribed to the NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) gene and starting 184bp away, we find three more genes of the Entner-Doudoroff Pathway, the first encoding a phosphogluconate dehydratase (EC 4.2.1.12) (contains Fe-S clusters [Outten FW, et al., 2004]) which is then overlapped by 46bp by the next gene, which encodes glucokinase (EC 2.7.1.2) (absolute requirement for divalent cation, in order of decreasing efficacy: Mn^{2+}>Mg^{2+}>Co^{2+}, [Klein, et al., 1986]) and 0bp after that, a gene encoding an oxidoreductase.

RegPredict investigation of this genome with regards to FUPA24, especially in conjunction with that of *H.ochraceum* yields several possible regulatory sequence predictions that may have a common orthologous regulatory region in *H.ochraceum*. The first of these is GtaAccgCcaATctGatcTatC (score=6.86) at position -246bp from the FUPA24 gene. Two others, CggaccCGaCggGgCGaccagG (score=6.33) and gCGGCGCgcaGCGCCGa (score=5.99) occur at -214bp and -98bp from the same gene. These are suggested by RegPredict to be related to a variably interpretted sequence found ~135bp upstream of the FUPA24 gene in *H.ochraceum*. The sequence is **GgcacCgCcGG**CCtGaGcgtgC (score=6.00), CgcGCCc**GgcaccgCcGG**CctG (score=6.46) or gCGcCc**GgcaCcGcCGg** (score=6.05), respective in comparison to the sequences described above (common sequence in **bold**). It is unclear whether weak but redundant regulatory sequences are at play at this locus in *H.chejuensis*, but several genes are predicted to be regulated with the FUPA24 gene in this organism, depending on which sequence is examined.

One such predicted operon is preceded by the sequence tgcGcCGgacGCcgtCGcCagg (score=5.12). The operon consists of 3 genes, the first of which encodes a perplasmic thiol:disulfide oxidoreductase DsbB, which is noted as being required for DsbA rexoxidation. This gene is followed 141bp later by one encoding a regulator of sigma D, which is then followed 147bp later by a gene encoding a FOG:WD40 repeat protein.

Another related sequence, CaGcCCCcaCagGgcGGGcCgG (score=5.11) precedes a single gene encoding phosphoenolpyruvate carboxykinase [ATP] (EC 4.1.1.49) (Mg^{2+} and either Ca^{2+} or Mn^{2+} [Krebs A, Bridger WA, 1976 & 1980 & Sudom A et al., 2003]).

The related sequence CgCacaGGgcgtcgCCgccGtG (score=5.10) precedes several genes encoding lactoylglutathione lyase, a probable deoxygenase, a transciptional regulator of the GntR family, a permease of the drug/metabolite transporter (DMT) superfamily and lastly 4-oxalocrotonate tautomerase (EC 5.3.2.-).

Yet another potential regulatory sequence, GGcgcAgCcGacCtGaTcgaCC (score=5.01) precedes a 4-gene operon which encodes porphobilinogen deaminase HemC (EC 2.5.1.61) (no known metal cofactor requirements), uroporphyrinogen-III synthase (EC 4.2.1.75) (no ion requirement observed [Alwan AF, et al., 1989]), a homologue of *E.coli* HemX protein and a homologue of *E.coli* HemY protein.

Lastly, the sequence aggGccCGaCGCGgCGacCagc (score=5.20) precedes a single gene encoding a protein described as a hemaglutinin in RegPredict, which is also noted as RTX toxins and related Ca^{2+} binding protein. This is the first hit in a NCBI BLAST as well. Other hits with very good scores (E values=0.0) are glycoside hydrolase family protein, cellulose 1,4 beta-cellobiosidase (cellulase) and exoglucanase.

The FUPA24 family's unique fused structure type and predominance in Actinobacteria (with the three exceptions, 2 of which are probably the result of horizontal gene transfer) are notable characteristics that may provide insight into their function. The unique lifestyle of the organisms most commonly found to possess the

FUPA24 gene may suggest that it is beneficial if not crucial to their survival. These genes are found to often be associated, either by colocalization or by putative coregulation, with enzymes with Mg^{2+} requirements. While the Type II topological portion of these transporters has been previously shown to be more closely related to Ca^{2+} (Family 2) transporters, such a major modification does open the possibility of highly modified specificity. Furthermore, the associations explored herein reveal that the FUPA24 gene may be expressed with biosynthesis genes of Coenzyme A, cytochrome c, and both lysyl- and isoleucyl- tRNAs. Perhaps access to, and allocation of, the divalent cations necessary for these processes requires special handling in such an environment. What effect does the deletion of this gene have on the fitness of a mycobacterium in its host environment?

The FUPA25 ATPase Family

FUPA25 is, generally speaking, a family of Type I P-type ATPases, however some lack either TMS A or both TMSs A and B (topological type VIII). The general topology of this family consists of 6 or 7 TMSs, depending on if only A or both A and B are absent. In any case, B and 1 & 2 cluster together within the range of residues 7-125, TMSs 3 & 4 cluster within residues 229-315 and 5 & 6 cluster with residues 553-641. The family has 11 members of sizes ranging from 645–776 aas in Actinobacteria (3.A.3.25.1, 3.A.3.25.4) (794aa, new since Chan, et al., 2010), 617-759 aas in proteobacteria (3.A.3.25.2), and 601-623 aas in firmicutes (3.A.3.25.3). This family is known to be most closely related to Family 6, and in a TCBLAST search, FUPA32 also comes up close to the top.

Actinobacteria:

TC 3.A.3.25.1

Blo2 (*Bifidobacterium longum* NCC2705 719aa, gi:23465139) is located amongst a variety of gene of known function, with 412bp of intergenic space upstream before a gene that is co-directional and 340bp of intergenic space downstream before a gene that is transcribed convergently.

The upstream gene encodes serine protease, DegP/HtrA, do-like (EC 3.4.21.-) and it is divergently transcribed from a gene encoding a tRNA-guanine transglycosylase

(EC 2.4.2.29) (predicted to have one Zn^{2+} per subunit by Uniprot). Downstream of this gene is another, larger intergenic space (1437bp).

Downstream of the FUPA25 gene but convergently transcribed to it, we find a AAA+ superfamily ATPase encoded, which is itself transcribed divergently to a gene 131bp away encoding ferredoxin--NADP(+) reductase. Actinobacterial (eukaryote-like) type [2Fe2S] (EC 1.18.1.2) is found next, followed 193bp later by a Zn-dependent protease with chaperone function, HtpX. The FUPA25 gene, ferredoxin reductase and htpX are conserved together in *Gardnerella vaginalis* 315-A as well.

RegPredict suggests several palindromic possible regulatory sequences preceding this FUPA25 gene, some of which overlap and have overlapping paralogues preceding other genes. For clarity, these sequences are preceded by a letter used to indicate which group of related sequences they belong to. One such overlapping pair of sequences is (*A)* *a***cgagCgTAaGgag**ca (score=6.40) and (*B) a*T**aCgagCGtaaGgAg** (score=6.40). Several related sequences are found elsewhere in the genome, such as a (*A)* *a***cgaCCgTAaGGag**ta (score=5.29) and (*B*) gg**aCgacCGtaaGgag** (score=5.15), which precede a gene encoding membrane alanine aminopeptidase N (EC 3.4.11.2) (contains Zn^{2+} [Ito K, et al., 2006]). Another related sequence set ((*A) a***CgagagTAaggaa**Ga, score=5.15 and (*B) a*g**aCgagagtaaGgaa**, score=5.53) precedes a gene encoding phosphoglucomutase (alpha-D-glucose phosphate specific) (EC 5.4.2.2) (requires Mg^{2+}, but Ni^{2+}, Mn^{2+} and Zn^{2+} at a concentration of 1 mM can also support activity to a lesser extent [Rashid N, et al., 2004]). Still another set of related overlapping sequences ((*A*) **tcgagCgTAaGgaa**cc, score=5.15 and (*B*) gg**tCgagCGtaaGgaa**, score=5.01) precedes a

gene encoding lactoglutathione lyase, which is noted as a glyoxalase/bleomycin resistance protein/dioxygenase. Lastly for sequence group A, a possible regulatory sequence ((*A) a*cCatCgTAgGgaGca, score=5.07) is found preceding a bicistron encoding hydroxyethylthiazole kinase (EC 2.7.1.50) (Mg^{2+} stabilizes the transition state and the phosphoryl group [Jeyakanthan J, et al., 2009]) and thiamin biosynthesis protein ThiC. Finally for sequence group B, RegPredict finds two sequences of questionable relatedness on their own, but one of which also has an overlapping sequence related to one found preceding the FUPA25 gene. The sequence (*B*) AAcaaCAC|GTGaacTT (score=5.05) precedes a gene encoding phosphoglycolate phosphatase (EC 3.1.3.18) (divalent cation required [Husic HD, et al., 1985]) and the sequence (*B*) **AAGaaggC|GtgacCT**T (score-5.78) precedes a gene encoding Zinc/Iron uptake regulation Zur/Fur Family protein. This second possible regulator overlaps with a second palindrome, which will be referred to as part of sequence group C, A**aagaAgG|CgTgaccT** (score=5.78) along with a third palindrome found to precede the FUPA25 gene (*C) (***TaacaAgG|**CgTgaccA, score=6.11). Finally, this third palindrome suggested to regulated the FUPA25 gene is found overlapping the fourth and last, which will be referred to as part of sequence group D, (*D*) CggTGggt|**taaCAagG** (score=6.72), which has a related sequence (CaaTGggt|gtaCAagG, score=5.28) preceding a bicistron encoding a two component transcriptional regulator of the winged helix family, VanR followed by a 454aa integral membrane sensor signal transduction histidine kinase with strong homology to putative antibiotic transporter sensor protein [*Clostridium botulinum*

NCTC 2916], length=459, E-value=8e-100, Identities=164/450 (36%), Positives=272/450 (60%), Gaps=(0%).

Cef8 (*Corynebacterium efficiens* YS-314 645aa, gi:25028496) is found encoded second in an operon which begins with a gene encoding a hypothetical protein with 42% similarity to LigA [*Nocardioidaceae bacterium* Broad-1] length=346, Expect=4e-05. The FUPA25 gene Cef8 is found 153bp later, followed another 29bp later by a gene encoding transcription factor accessory protein with a S1 RNA-binding domain. After another 39bp, a radical SAM domain heme biosynthesis protein is encoded.

Convergent to this is a 5-gene thiamin biosynthesis operon, transcribed in the order ThiEOSGF. ThiE, thiamin-phosphate pyrophosphorylase (EC 2.5.1.3) requires one of the divalent cations of Mg, Mn, Ca, Co or Zn. The magnesium and manganese ions are equally effective, the others 20% as much so (Kawasaki T, et al., 1973 & 1979).

Divergently transcribed from the operon containing the FUPA25 gene, 87bp away there is a transposase subunit A encoded. The next gene downstream of this, which is co-directional, encodes a hypothetical protein with homology to putative 5,10-methylene- tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase [*Saccharopolyspora erythraea* NRRL 2338], length=376, Expect=1e-39, positives=(51%). This gene is probably not involved with the FUPA25 gene, and downstream of it are various fragments of genes of transposases encoding 67-301aa proteins.

RegPredict suggests two overlapping regulatory sequences, centered 10bp apart, **CtTgTtGca**tCgAaAgG (score=7.14) and TGgcACg**cttGTtgCA** (score=6.93), upstream

of the operon containing the FUPA25 gene. Similar predicted regulatory sequences are found ahead of several other genes as well. The sequence CtTgTttccacgAaAgG (score=5.46) is found preceding a DUF1541 domain-containing protein with homology to the putative lipoprotein YdhK [*C.amycolatum* SK46], length=226, E-value=2e-69, Identities=109/214 (51%), Positives=137/214 (64%). The sequence CCTgTtgCaGagAaAGG (score=5.46) is found preceding a 6-gene operon encoding a maltose/maltodextrin ABC transporter/substrate binding periplasmic protein, MalE, followed 164bp later by a 161aa hypothetical protein homologous to Activator of Hsp90 ATPase 1 family protein [*Tsukamurella paurometabola* DSM 20162], length=155, E-value=8e-44, Identities=68/149 (46%), Positives=96/149 (64%). Following this 149 bp later is encoded a N-acetyl-D-glucosamine ABC transport system permease, whose gene is then followed by a 3bp overlapping gene encoding a Maltose/maltodextrin ABC transporter permease, MalG. Another 11bp after this is encoded a 77aa hypothetical protein whose only significant homology is to putative secreted protein [*C.ulcerans* 809], length=67, E-value=6e-21, Identities=37/66 (56%), Positives=54/66 (82%). This is followed 7bp later by a gene encoding a bifunctional deaminase-reductase domain protein. The sequence TggcaCgctaGctgaA (score=5.20) precedes a bicistron encoding a 203aa hypothetical protein with homology to vitamin K epoxide reductase [*Brevibacterium linens* BL2], length=208, E-value=4e-62, Identities=90/191 (47%), Positives=133/191 (70%), followed 154bp later by a gene encoding a ribonucleotide reductase of class IB (aerobic), beta subunit (EC1.17.4.1) (requires Mn^{2+} [Ogata H, et al., 2009]). Lastly, two operons are found preceded by identical sequence at the loci CE1228

and CE1465. The former is a 4-gene operon encoding first an insertion element fragment 103aa in length, but found with homologues in transposases of several times this size, then second another putative transposase 125aa long 51bp later, another one, 102aa long, 25bp later which overlaps a fourth transposase fragment gene, encoding a 298aa transposase. The latter cistron consists of 3 genes encoding putative transposases 447, 103 and 285aa long, 60 and 51bp apart. Using RegPredict to view the paralogues and orthologues of genes in a given operon, it becomes apparent that this genome is riddled with transposase fragments.

Cgl3 (*C.glutamicum* ATCC 13032 650aa, gi:19551725) is encoded by the fourth gene in a cluster of genes transcribed codirectionally. However, neither BioCyc nor RegPredict indicate that it is included in an operon with them, but rather that it is the at the beginning of a 3-gene operon of its own. The operon upstream of it, from beginning to end encodes GABA:alpha-ketoglutarate aminotransferase (EC 2.6.1.19) (no iron-sulfur cluster [Liu W, et al., 2004]), followed 2bp later by succinate-semialdehyde dehydrogenase [NAD(P)+] (EC 1.2.1.16) (bivalent cations such as Mg^{2+}, Mn^{2+}, Ca^{2+} or Fe^{2+} are not required [Sanchez M, et al., 1989]) and then 39bp after that is encoded an amino acid permease. The FUPA25 gene is found 126bp later, followed 19bp later by a gene encoding a protein referred to as phytoene dehydrogenase and related proteins, which is then followed by a gene encoding a 202aa SIMPL superfamily hypothetical protein. Another 299bp downstream and codirectionally transcribed, genes encoding LSU ribosomal protein L10p (P0) and LSU ribosomal protein L7/L12 (P1/P2) are found,

themselves 78bp apart. Another 526bp later is found encoded a 336aa hypothetical membrane protein with no significant homology and two predicted TMSs, one toward each terminus. After another 421bp, more members of an apparent ribosomal subunit cluster are encoded transcribed as follows:

1) DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)
2) 84bp later, DNA-directed RNA polymerase beta` subunit (EC 2.7.7.6)
3) 176bp later, 61aa hypothetical membrane protein with no significant homology and 2 TMSs
4) 12bp later, XRE family transcriptional regulator
5) 29bp later, a 232aa hypothetical protein is an exception, as it is convergently transcribed, with closest informative homology to LigA [*Nocardioidaceae bacterium* Broad-1], length=222, E-value=2e-20, Identities=67/212 (32%), Positives=105/212 (50%), Gaps=28/212 (13%)
6) 323bp upstream of its start site, SSU ribosomal proteins S12p (S23e)
7) 7bp later, S7p (S5e)
8) 348bp later, the translation elongation factors G and
9) 363bp later, translation elongation factors Tu,

Further on are more ribosomal subunit genes.

RegPredict suggests several possible regulatory sequences found upstream of the FUPA25 gene. One of these is gTcgTtgGcCttAttAt (score=7.14). Two similar sites are found within the genome that fit with this sequence, they are tTcgTtgGcCtgAtcAt (score=5.46) and gTAgTAgTcCtTAtTAa (score=5.46), and they respectively precede a

tricistron encoding cobalamin/Fe^{3+}-sideophores transport system, ATPase component, then a 319aa hypothetical protein that appears to be an elongated homologue of putative heme transporter HtaA [*C.glutamicum*], Length=186, Score=338 bits (867), E-value=9e-116, Identities=165/187 (88%), Positives=176/187 (94%), Gaps=1/187 (1%). The last gene in this operon is a 279aa hypothetical protein with similar homology to the same protein (Score=186 bits (473), E-value=1e-56, Identities=97/186 (52%), Positives=123/186 (66%), Gaps=8/186 (4%)). The second sequence precedes a lone coding region encoding a 737aa hypothetical protein with substantial homology to KAP family P-loop domain protein [*Arthrobacter phenanthrenivorans* Sphe3], Length=702, Score=166 bits (421), E-value=1e-40, Identities=129/437 (30%), Positives=209/437 (48%), Gaps=30/437 (7%).

Another predicted regulatory sequence for the FUPA25 gene is AgGGgtcgttggCCtT (score=6.93). The related sequence AgGtgtggttcgcCtT is found preceding an 8-gene operon encoding the following F0F1 ATP synthase subunit proteins in order of transcription:

1) subunit A, atpB
2) subunit C, atpE
3) subunit B, atpF
4) subunit delta, atpH
5) subunit alpha, atpA
6) subunit gamma, atpG
7) subunit beta, atpD

8) subunit epsilon, atpC

The entire F0F1 ATP synthase has the EC 3.6.3.14 and is thus indicated as a H^+-transporting two-sector ATPase according to BRENDA, and a divalent cation is required, in the order of decreasing efficiency: Mg, Mn, Co, Zn, Ni, Ca (Satoh S, et al., 1994).

Sco1 (*Streptomyces coelicolor* A3(2) 776aa, gi:21218721) is found encoded within a gene cluster which is conserved to varying degrees in a variety of other Actinomyceteae. The FUPA25 itself is probably more likely to transport Heavy Metals (Family 6) rather than Cu^{2+} (Family 5), and is suggested to with high affinity and low efficiency, based both on the general schema of P-type ATPases as well as another cluster member's. The FUPA25 gene has two codirectionally transcribed genes upstream of it, the closer one (123bp away) encoding a putative gluconate kinase and one 178bp upstream of that encoding a nitroreductase. Divergently transcribed from this, 370bp away, is a gene encoding helix-turn-helix domain DNA-binding protein.

Downstream of the FUPA25 gene, we find a convergently transcribed gene, which overlaps the FUPA25 gene by 21bp, encoding a 137aa hypothetical protein with homology to ABC transporter, ATP-binding protein [*Streptomyces griseoflavus* Tu4000], length=131, Score=171 bits (432), E-value=2e-53, Identities=90/128 (70%), Positives=99/128 (77%), Gaps=1/128 (1%). The next gene downstream from the FUPA25 gene is transcribed divergently from this (in the same direction as the FUPA25 gene), 218bp away. It encodes a putative regulator highly homologous to RNA

polymerase sigma factor, putative [*Saccharopolyspora erythraea* NRRL 2338], length=286, Score=382 bits (982), E-value=1e-132, Identities=184/255 (72%), Positives=207/255 (81%), Gaps=0/255 (0%). Convergent to this gene, 48bp away, is encoded a member of the universal stress protein family, and 659bp away from and divergent from that is encoded a regulator protein highly homologous to cyclic nucleotide-binding domain-containing protein [*Streptomyces sp. e14*], length=189, Score=195 bits (495), Expect=3e-61, Identities=103/177 (58%), Positives=117/177 (66%), Gaps=1/177 (1%). After another 68bp there is first encoded in the same direction one CBS domain protein, and a second one convergently encoded to the first, 331bp away. Lastly, transcribed divergently from this, there is a gene encoding a nicotinate phosphoribosyltransferase (EC 2.4.2.11). Homologues to all but the ABC transporter, putative RNA polymerase sigma factor and cyclic nucleotide-binding domain-containing protein homologue are found to be encoded near the FUPA25 and also each other in the following organisms: *Mycobacterium sp. MCS, Pseudonocardia dioxanivorans* CB1190, *Saccharopolyspora erythraea* NRRL 2338, *Streptosporangium roseum* DSM 43021, *Thermobispora bispora* DSM 43833, *Micromonospora aurantiaca* ATCC 27029, *Salinispora tropica* CNB-440 and *Verrucosispora maris AB-18-032*.

RegPredict suggests numerous conserved palindromic potential regulatory sequences, many of which are conserved with the genes they are implicated to regulate in *S.griseus* and *Mycobacterium sp*. MCS as well. One such sequence and those related to it in *S.coelicolor,* as well its orthologues in *S.griseus* and *Mycobacterium sp*. MCS, are described here. The first found preceding the FUPA25 gene is GGTcggcgtcgacACC

(score=6.27), which has a related sequence in *M.sp*. MCS (GGTGCgtgacgGCACC, score=5.64) preceding its FUPA25 gene. Another related pair of sequences in *S.coelicolor* (GGTcgtcgaaggcACC, score=5.13) and *M.sp* MCS (GGtGggcGCcgaCcCC, score=5.00) precedes a gene encoding Mannose-1-phosphate guanylyltransferase (EC 2.7.7.13) (Requires Mg^{2+} (best) or Co^{2+} [Ning B, et al., 1999]), although in *M.sp* MCS there is also a glycosyl transferase family 2 protein encoded 0bp upstream. Another sequence in *S.coelicolor* (GgTcGacg|acgCcAgC, score=5.00) precedes a tricistron encoding a NADH-ubiquinone oxidoreductase chain L, then chains M and N (EC 1.6.5.3). The holoenzyme may contain as many as 9 Fe-S clusters [Flemming D, et al., 2006]. This very same sequence is also found to precede this operon in *S.griseus*, with an additional gene encoding an ATP-dependent DNA helicase, RecQ at the end. These genes are also found regulated together in *M.sp* MCS, but Chain N is found separated from chains L & M, which are still together. The operons containing them are preceded by GgTGCgcg|acgGCAaC (score 5.25) and GgTcGacg|acgCcAaC (score=5.00), respectively. Additionally, a related 4-gene cistron is found in *S.coelicolor*, which seems to be unique to it. It is preceded by the sequence GGtcgtcg|tcgacgCC (score=5.13) and encodes NADH-ubiquinone oxidoreductase chains KLMN, though only chain N exhibits very strong homology to those previously mentioned.

Another sequence (GGTGgtcg|tcgaCACC, score=5.52 in *S.coelicolor* and GGTcgtcg|tcgacACC, score=5.73 *S.griseus*) is found preceding a gene encoding the alpha-amylase Malto-oligosyltrehalose synthase (EC 5.4.99.15) (highly active in the

presence of Ca^{2+} and Zn^{2+}, partially inhibited by Cu^{2+}, Hg^{2+}, Mg^{2+} and Mn^{2+} [Wu S, et al., 2010])

One probable set of sequences is of special interest because they are found in tandem in two streptomycetes. The sequences (GcTGCgcgacgGCAcC, score=5.25 and GGTcgCCgaGGacACC, score=5.13 in *S.coelicolor*; GGTcgCCgaGGacACC, score=5.13 and GcTcCgcgacgGcAcC, score=5.46 in *S.griseus*) precede homologous 9-gene operons in both operon, the genes encoded in *S.coelicolor* will be described in detail, in order of transcription:

1) Trp region conserved putative membrane protein
2) 145bp later, an 84aa putative membrane protein with 2 predicted TMSs and no significant homology
3) 103bp later, a 148aa putative membrane protein with 3 predicted TMSs and no significant homology
4) 118bp later, indole-3-glycerol-phosphate synthase trpC (EC 4.1.1.48) (Ca^{2+}, Mg^{2+}, Mn^{2+} and Na^{+} "significantly affects activity, the optimal concentration is about 0.4-2.0 mM" (Yang Y, et al., 2006)
5) 21bp later, 63aa hypothetical protein with 0 predicted TMSs and no significant homology
6) 115bp later, Tryptophan synthase beta chain trpB (EC 4.2.1.20)
7) -3bp later (overlap), Tryptophan synthase alpha chain trpA (EC 4.2.1.20)

8) 48bp later, 276aa putative 1 TMS integral membrane protein with homology to DSBA oxidoreductase [*Streptomyces sp*. e14], length=256, Expect=4e-137, Identities=192/257 (75%), Positives=221/257 (86%), Gaps=2/257 (1%)
9) 77bp later, prolipoprotein diacylglyceryl transferase (EC 2.4.99.-)

The related sequence set with the oldest known divergence found here is GGTcgCcGCcGgcACC (score=5.27) in *S.coelicolor* with the related sequences GGTcgCCGCGGgcACC (score=4.67) in *M.sp.* MCS and GGTGgCcGCcGgCACC (score=5.06) in *B.longum*. In all cases the sequences precede a gene encoding acetolactate synthase, small subunit ilvN (EC 2.2.1.6). The enzyme is active in the presence of Mn^{2+}, Mg^{2+}, Ca^{2+}, Cd^{2+}, Co^{2+}, Zn^{2+}, Cu^{2+}, Al^{3+}, Ba^{2+} or Ni^{2+}, the activity is about 50% for Ni^{2+} and 133% for Mn^{2+} as compared to Mg^{2+} (Duggleby RG, et al., 2008). In *B.longum* two genes encoding plasmid stability proteins follow this, whereas in the others it is followed by one encoding ketol-acid reductoisomerase ilvC (EC 1.1.1.86) (Mg^{2+} required; Mn^{2+}, Co^{2+}, Ni^{2+}, Zn^{2+}, Ca^{2+}, Cu^{2+} and Co^{3+} can not substitute for Mg^{2+} [Dumas R, et al., 2001]).

TC 3.A.3.25.4

This P-type ATPase family includes a C-terminal hemeerythrin (Hr) domain (Traverso *et al.*, 2010). This domain binds two iron ions per monomer (a diiron center) and is thought to regulate or provide more direct function in iron transport (Traverso *et al.*, 2010).

Acel (*Acidothermus cellulolyticus* strain ATCC 43068 / 11B; 794aa, gi:117927237) is encoded in by a gene with a 597bp intergenic space preceding it, and thus is predicted to be the first transcribed gene of any operon to which it may belong. The gene upstream of it, which is codirectionally transcribed, encodes a putative lipoprotein, as is the next one upstream (192bp away), encoding a putative proteasome assembly chaperone 2. The next gene upstream of that is divergently transcribed (235bp away) and encodes an aerobic glycerol-3-phosphate dehydrogenase (EC 1.1.5.3) (activated by Ca^{2+} and contains 2 mol of non-heme iron per dimer [Schryvers A, et al., 1978 & 1981]). Downstream of the FUPA25 gene is an arrangement of genes that is likely to form an operon. The predicted operon encodes as follows, in order of transcription:

1) FUPA25
2) kelch repeat-containing protein
3) tRNA-Leu-CAG
4) forkhead associated (FHA) domain containing proteins, putative nuclear signaling proteins which may bind phospho-threonine, -serine and possibly –tyrosine
5) another FHA
6) putative protein phosphatase
7) cell division protein FtsW

8) cell division protein FtsI [peptidoglycan synthetase] (EC 2.4.1.129) (slight stimulation with Mg^{2+}, but no divalent cation required for function [Nakagawa J, et al., 1984])
9) serine/threonine protein kinase PrkC, regulator of stationary phase

RegPredict suggests several potential palindromic regulatory sequences preceding the FUPA25 gene, the first of which is tGcgagtCGgacacCg (score=6.93). Four related sequences were found. The first related sequence gGcgGatCGgaCacCg (score=5.20) precedes a bicistron encoding a methionyl-tRNA sythetase (EC 6.1.1.10) (this enzyme most likely lacks a Zn^{2+} binding motif and contains no Zn^{2+} and is absolutely dependent on Mg^{2+} [Deobagkar DN, et al., 1976; Kim S, et al., 1998]). This gene is overlapped by 3bp by a gene encoding a putative deoxyribonuclease YcfH. The second related sequence is also found (tgcgGgtCGgaCgcgg, score=5.20) preceding a gene encoding an aerobic glycerol-3-phosphate dehydrogenase (EC 1.1.5.3). This is the same dehydrogenase described above as being distantly divergently transcribed from the FUPA25 gene, however the sites are most certainly unique. The third related sequence (CGccagtctgacacCG, score=5.20) precedes a bicistron encoding a transcriptional regulator of the MarR family, followed with a 3bp overlap by a gene encoding a multidrug resistance protein B. The fourth related sequence (gGagactCGgacacCg, score=5.20) precedes a 5-gene operon encoding a metal dependent phosphohydrolase with a HD-GYP domain (putative Zn^{2+} and Mg^{2+} binding motifs), then a 426aa protein described as a putative lipoprotein, but with strong homology to both metal dependent phosphohydrolase ([*Thermobispora bispora* DSM 43833], length=427, Expect=9e-95,

Identities=179/408 (44%), Positives=247/408 (61%), Gaps=37/408 (9%)) as well as response regulator receiver protein ([*Streptosporangium roseum* DSM 43021], length=429, Expect=9e-86, Identities=179/410 (44%), Positives=234/410 (57%), Gaps=40/410 (10%)), then a methylcrotonyl-CoA carboxylase biotin-containing subunit (EC 6.4.1.4), then a biotin carboxyl carrier protein and finally a twitching motility protein, PilT.

The next sequence predicted by RegPredict as a palindromic regulatory sequence preceding the FUPA25 gene is g**cgttGtcagCgcacg** (score=6.93). Three related sequences are described, the first of which (gcgtaGtgatCgcacg, score=5.20) precedes a biscistron encoding aspartate/tyrosine/aromatic aminotransferase class I and II, followed by a fructokinase (EC 2.7.1.4) (highest activation by free Mg^{2+}, Co^{2+} can partially replace Mg^{2+} in activation, but no activation by Cd^{2+}, Cu^{2+} or Mn^{2+} [Sabater B, et al., 1972]). The second related sequence (gcgTtGTg|aACgAacg, score=5.20) precedes a bicistron encoding a GTP-binding protein Obg, overlapped by 3bp by a gene encoding a glutamate 5-kinase (EC 2.7.2.11) (forms dimers requiring two Mg^{2+} each [Marco-Marin C, et al., 2007]). The third related sequence (gcgtCGCC|GGCGcacg, score=5.20) precedes a single gene encoding a 177aa hypothetical integral membrane protein with 4 TMSs but no significant homology.

The next sequence predicted by RegPredict as a palindromic regulatory sequence preceding the FUPA25 gene is **CgtTGtca|gcgCAcg**G (score=6.93), which almost entirely overlaps the previously described one (overlap em**bold**ened). Two sets of genes are preceded by related sequences, the first of which is preceded by the same related

sequence in duplicate. The sequence (CgaTGccagctCAcgG, score=5.20) is found 182bp and 243bp upstream of a 3-gene operon encoding a (R)-citramalate synthase (EC 2.3.1.182) (Mg^{2+} and Mn^{2+} are the best activators [de Carvalho LP, et al., 2006]), which is a 2-isopropylmalate synthase/homocitrate synthase family protein, then a 5-carboxymethyl-2-hydroxymuconate Delta-isomerase (EC 5.3.3.10) (no known metal cofactors) and finally a glutamyl-tRNA synthetase (EC 6.1.1.17) (requires Zn^{2+} and Mg^{2+} [Liu J, et al., 1993; Saha R, et al., 2009]). The other related sequence (CgTTGgaa|gcgCAAgG, score=5.20) precedes a 3-gene operon encoding a transcriptional repressor of the fructose operon, DeoR family, the a DhnA-type fructose-1,6-bisphosphate aldolase and lastly a glucokinase (EC 2.7.1.12) (activated by Mg^{2+} [Cohen SS, 1951]).

The next sequence predicted by RegPredict as a palindromic regulatory sequence preceding the FUPA25 gene is GcTcaCTt|tAGccAtC (score=6.93). Two related sequences are described the first of which (GcTGGcgt|cgtCCAcC, score=5.20) precedes a bicistron encoding ribonucleotide reductase transcriptional regulator NrdR followed by a ribonucleotide reductase of class II (coenzyme B12-dependent) (EC 1.17.4.1) (binds Mn(II)/Fe(II) and Mn(III)/Fe(III) clusters in separate sites [Pierce BS, et al., 2005]). The second related sequence (GcTGacGCGCtgCAcC, score=5.20) precedes a gene encoding a single small regulatory protein, RecX.

The fifth and last related site preceding the FUPA25 gene described here (ctCaccTgaAacgGtc, score=6.93) has a related sequence in cgCacctGCgacgGtc (score=5.20), which precedes a gene encoding a WhiB-type transcriptional regulator.

Proteobacteria

TC 3.A.3.25.2 (7 TMSs and an extra putative N-terminal TMS):

Mlo7 (*Mesorhizobium loti* MAFF303099 (α) 617aa, gi:13475503) is encoded at the beginning of a 5-gene operon which most likely is expressed together with another gene which is divergently transcribed. The genes following in the direction of transcription beginning with the FUPA25 gene encode:

1) FUPA 25
2) 28bp later, thioredoxin
3) -36bp later (overlap), MgtC family protein
4) 86bp later, 129aa hypothetical protein with no significant homology and no TMSs
5) 33bp later, a putative NADPH-quinone reductase (modulator of drug activity B)

The divergent gene 167bp away encodes an acetyl-CoA synthetase (ADP-forming), alpha and beta chains. This gene is presumed monocistronic because it is met head-to-head with another encoding a putative gluconate kinase. The acetyl-CoA synthetase gene is found to consistently lie near one encoding a universal stress protein, UspA and related nucleotide-binding protein across many genomes. In *Chelativorans sp.* BNC1 and *Sinorhizobium meliloti* 1021, these two cluster near genes encoding a Family 2 (Ca^{2+}) P-type ATPase. In *Sinorhizobium meliloti* 1021, this is because the Family 2 gene is 4.6kb upstream from the FUPA25 gene. In *Chelativorans sp.* BNC1 the Family 2

gene is 17.6kb upstream from the FUPA25 gene and the FUPA25 gene is 4.6kb upstream from the CcoNOQPGH-27(I)-S operon (see FUPA27).

For *M.loti*, RegPredict suggests that the sequence preceding the FUPA25 of ctGtTcATgGCgATaAtCgc (score=7.75) may be regulatory in nature. In *M.loti*, as well as *S.meliloti*, related sequences occur (CCGgTcATgacgATcAtCGG, score=5.16 in *M.loti*, CCGgTcATggtgATcAtCGG, score=5.16 in *S.meliloti*) which in both species are suggested to regulate similar operons. In *M.loti*, this operon encodes (in order of transcription):

1) conserved hypothetical protein probably involved in sulfate reduction
2) 46bp later, sulfite reductase [NADPH] hemoprotein beta-component (EC 1.8.1.2) (Fe4S4 cluster, McRee DE, et al., 1986)
3) -25bp (overlap), phosphoadenylyl-sulfate reductase [thioredoxin] [cysH] 1.8.4.8
4) 112bp later, oxidoreductase probably involved in sulfite reduction
5) 96bp later, alkan-1-ol dehydrogenase, PQQ-dependent (EC 1.1.99.20) (no cation requirement [Kawai F, et al., 1980])
6) ThiJ/PfpI family protein

   In the *S.meliloti* operon, 3 of these are missing (#s 3, 5 and 6) and after the oxidoreductase gene there is is instead a gene 146bp later encoding a ferredoxin--NADP(+) reductase (EC 1.18.1.2)

Another related sequence suggested by RegPredict (ctGcacaTgccgAgaatCgt, score=5.16) precedes a 5-gene operon encoding the following in order of transcription:

1) tRNA-i(6)A37 methylthiotransferase
2) 0bp later, phosphate starvation-inducible ATPase PhoH with RNA binding motif
3) 19bp later, a 166aa hypothetical protein with no known TMSs but homology to diacyglycerol kinase [*Brucella melitensis bv. 1 str.* 16M], length=168, Score=202 bits (514), Expect=8e-65

   Identities=101/154 (66%), Positives=118/154 (77%), Gaps=1/154 (1%), though it is suggested to be a putative metal-dependent hydrolase, as a member of the UPF0054 superfamily
4) 28bp later, Magnesium and cobalt efflux protein CorC
5) glyoxalase family protein

Lastly for this sequence set, the sequence ctGtTCggCGCGatGAtCgt (score=5.16) precedes 5 genes in a 6-gene operon encoding the following (in order of transcription):

1) fructose ABC transporter, ATP-binding component, FrcA
2) 85bp later, glycogen phosphorylase (EC 2.4.1.1) (no divalent cations required [Robson RL, et al., 1974])
3) 107bp later, 1,4-alpha-glucan (glycogen) branching enzyme, GH-14-type (EC 2.4.1.18) (Mg^{2+} 15% increased enzyme activity, 5 mM [Yoon SA, et al., 2008])
4) 54bp later, glucose-1-phosphate adenylyltransferase (EC 2.7.7.27) (Mg^{2+} required [Preiss J, 1978])

5) 84bp later, glycogen synthase, ADP-glucose transglucosylase (EC 2.4.1.21) (no known cation requirements in bacteria)
6) phosphoglucomutase (EC 5.4.2.2) (requires Mg^{2+}, but Ni^{2+}, Mn^{2+} and Zn^{2+} at a concentration of 1 mM can also support activity to a lesser extent [Rashid N, et al., 2004])

Another suggested regulatory sequence preceding the FUPA25 gene of *M.loti* is tgtctCcTGttCAtGgcgat (score=7.75). It has many related sequences as suggested by RegPredict, such as tgTctCgatttcatGccAat (score=5.16), which precedes a 6-gene operon encoding the following (in order of transcription):

1) N-(5'-phosphoribosyl) anthranilate isomerase (EC 5.3.1.24) (no known cation requirements)
2) 89bp later, Tryptophan synthase beta chain trpB (EC 4.2.1.20)
3) 63bp later, Tryptophan synthase alpha chain trpA (EC 4.2.1.20)
4) 49bp later, Acetyl-CoA carboxylase carboxyl transferase, beta subunit (EC 6.4.1.2) (the enzyme requires Mg^{2+} or Mn^{2+} for coordinating the ATP phosphates for catalysis, possibly Zn^{2+} [Benson BK, et al., 2008])
5) 102bp later, Folylpolyglutamate synthetase (EC 6.3.2.17) (K^{+} and Mg^{2+} required [Bognar AL, et al., 1986])
6) 28bp later, 140aa hypothetical protein with no TMSs and homology to Glyoxalase/bleomycin resistance protein/dioxygenase [*M.opportunistum*

WSM2075] length=139, Score=263 bits (672), Expect=9e-90, Identities=121/139 (87%), Positives=132/139 (95%), Gaps=0/139 (0%)

Another related sequence (tgtGtCgtGtgCctGcCgat, score=5.16) precedes a 3-gene operon which encodes an ABC transporter protein, ATP binding component/ monosaccharide-transporting ATPase (EC 3.6.3.17) (no known cation requirements), followed 82bp later by a ABC transporter, permease protein gene and lastly 79bp later an ABC transporter binding protein.

Four other related sequences preceded single genes. The score is the same for all of them (score=5.16). The sequences and encoded proteins are:

a) tgtGtCctgctgttGcCgat for glutathione synthetase (EC 6.3.2.3) (divalent metal ion required, Mg^{2+} is most effective, Co^{2+} and Mn^{2+} can substitute with weaker activity [Gushima H, et al., 1983])
b) tcactCgctt|tcatGgcgat for transcriptional regulator, AsnC family
c) tgtttCcTgc|ttAtGgcgtc for 3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11) (Mg^{2+} activates is required for activity and is most active, Mn^{2+}, Ni^{2+}, Co^{2+} and Zn^{2+} are progressively less active [Powers SG, et al., 1976 & 1979])
d) cgttgCcTGt|tCAtGgccac for a 324aa hypothetical protein with 1 TMS at the N-terminus and homology to putative carboxylesterase [*Ahrensia sp*. R2A130] length=322, Expect=8e-105, Identities=152/322 (47%), Positives=209/322

(65%), Gaps=0/322 (0%)

Sme9 (*Sinorhizobium meliloti* 1021 (α) 746aa, gi:16263085) is encoded at the beginning of a 5-gene operon which most likely is expressed together with another operon which is divergently transcribed. The genes following in the direction of transcription beginning with the FUPA25 gene encode:

1) FUPA 25
2) small molecule metabolism/putative hydrolase protein
3) Putative NADPH-quinone reductase (modulator of drug activity B) (See *M.loti*)
4) 114aa hypothetical protein with no TMSs or significant homology
5) 104aa hypothetical protein with 0 TMSs and homology to possible CycB2 cytochrome c552 precursor [*Rhodopseudomonas palustris* CGA009], length=104, Expect=7e-20, Identities=42/92 (46%), Positives=56/92 (61%), Gaps=1/92 (1%)

The divergent gene 115bp away encodes a 284aa, 8 TMS putative transmembrane protein. Another 152bp later, The N-terminus of a MgtC family protein (see *M.loti*) is encoded. While two genes follow this one, each overlapping the previous (by 3bp and 52bp), it is highly likely due to homology that this is actually one MgtC protein-encoding gene, or a functional divided one. After these, 109bp later, a member of the UspA family is encoded, then 110bp after that, Alcohol dehydrogenase, Zn-

dependent class III. Convergently encoded to this gene, 66bp away is encoded a Family 2 P-type ATPase (Ca^{2+} transporter) as mentioned previously.

Regpredict suggests a variety of possible palindromic regulatory sequences preceding the FUPA25 gene, most of which are found ~60bp ahead of the gene. The overlap of these sequences is denoted by **bold**face. The first of these is **tGCGaacgacgc**CGCg (score=5.03) and the sequence related to it (tgCGcCcggcGcCGgg, score=5.03) precedes a 4-gene operon encoding the following proteins (in order of transcription):

1) probable sugar ABC transporter, permease protein
2) 2bp later, sugar ABC transporter, permease protein
3) 197bp later, putative ABC transporter, ATP-binding protein/Glycerol-3-phosphate-transporting ATPase
4) 0bp later, probable ABC transporter, ATP-binding protein/Glycerol-3-phosphate-transporting ATPase

The next suggested regulatory sequence palindrome preceding the FUPA25 gene is GCGc**tGcgaaCgaCGC** (score=5.31). The related sequence GCGCaGGgaCCgGCGC (score=5.12) is found preceding another 4-gene operon encoding the following proteins (in order of transcription):

1) putative sugar uptake ABC transporter periplasmic solute-binding protein precursor/LacI transcriptional regulator
2) -3bp later, ribose/sugar uptake ABC transporter ATP-binding protein

3) -7bp later, putative sugar uptake ABC transporter permease protein
4) 3bp later, esterase or acylase, CocE/NonD family protein

The next suggested regulatory sequence palindromes overlap each other preceding both the FUPA25 gene and the suggested coregulated operon (underlined). The first is CGC**tGCGaacgaCGC**cGCG (score=5.70) and the second is GatgCGc**TgcgAaCGacgC** (score=5.64). The almost entirely overlapping related sequences (respectively) cGaaGcggatcgaaCgaCa (score=5.47) and gaagCGgatcgaaCGacat (score=5.02) are found preceding a 2-gene operon encoding the following proteins (in order of transcription):

1) ABC-type xylose transport system, permease component (XylH)
2) 15bp later, D-xylose transport ATP-binding protein XylG

The next (and last overlapping) suggested regulatory sequence palindrome preceding the FUPA25 gene is GCGc**tGcgaacgaCgcCGC** (score=5.74). The related sequence GCGcaccgGcCgtcgcCGC (score=5.10) is found preceding a lone gene encoding pyridoxal 4-dehydrogenase Pld/ putative L-fucose-beta-pyranose dehydrogenase.

Another suggested regulatory sequence (tGatccaagTgA**acccccgCc**, score=5.09) preceding the FUPA25 gene has a related sequence (tGgGagaGcTcAaCgacCgCc, score=5.14) preceding a 2-gene operon encoding first a 194aa hypothetical lipoprotein

with no apparent TMSs and no significant homology and then an isoprenylcysteine carboxyl methyltransferase.

Another suggested regulatory sequence (with overlap to the one above in **bold**) (**acCcccgcC**caGtaaaaGca, score=5.44) preceding the FUPA25 gene has a related sequence (cGCaccGCCCGGGCacaGCa, score=5.45) preceding a bicistron encoding a glutamate racemase required for biosynthesis of D-glutatmate and peptidoglycan protein (EC 5.1.1.3) (no known metal/ion requirements) followed by an AAA ATPase.

The last suggested regulatory sequence described here (ccggTGCggagttGCAagct, score=5.22) preceding the FUPA25 gene has a related sequence (acCgcGCgGagCgGCagGcc, score=5.10) preceding a lone gene encoding a SMP-30/gluconolaconase/LRE domain-containing protein (EC 3.1.1.17) ("the purified enzyme is strictly dependent on Ca^{2+} and undergoes rapid denaturing precipitation on Ca^{2+} depletion even in the presence of detergent" [Shinagawa E, et al., 2009]).

Tcr4 (*Thiomicrospira crunogena* XCL-2 (γ) 759aa, gi:78485067) is encoded at the end of an exceptionally long gene sequence, well within a reasonable distance to be the last or second to last gene in what may be an extensive operon. However, it should be noted that this FUPA25 gene is only found with the other genes in this suggested operon in *T.crunogena*, and that there is no known support for this relationship elsewhere. It is worth noting as well that, as mentioned previously, the first and second most frequently co-localized genes to the FUPA25 genes are those encoding UspA and P-type ATPases, most often for Ca^{2+} or Mg^{2+}, but also one that appears to be a FUPA30 and another that

appears to be a FUPA24. *Mycobacterium parascrofulaceum* specifically has 4 P-type ATPases (Ca^{2+}, Mg^{2+}, FUPA24, FUPA25) clustered very close together.

The gene series that this FUPA25 gene is suggested to be a part of encodes the following, in order of transcription:

1) Ribosomal large subunit pseudouridine synthase C (EC 4.2.1.70) (dependent on presence of Mg^{2+}, Co^{2+}, Fe^{2+} or Mn^{2+}, inhibited by Zn^{2+} and Ni^{2+} [Heinrikson RL, et al., 1964; Preumont A, et al., 2008])
2) 41bp later, a protein similar to phosphoglycolate phosphatase, which is conserved clustered with ribosomal large subunit pseudouridine synthase C
3) Maf/YceF/YhdE family protein (opposite direction of the rest of the series)
4) COG1399 protein, clustered with ribosomal protein L32p
5) LSU ribosomal protein L32p
6) Phosphate:acyl-ACP acyltransferase PlsX
7) 76bp later, 3-oxoacyl-[acyl-carrier-protein] synthase, KASIII (EC 2.3.1.41) (stabilized by Mg^{2+} [Price AC, et al., 2003])
8) 87bp later, Malonyl CoA-acyl carrier protein transacylase (EC 2.3.1.3) (no known metal ion required)
9) 100bp later, 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) (binds NADP+, does not use metal cofactor [Silva RG, et al., 2006 & 2008])
10) 712bp later, Acyl carrier protein
11) 220bp later, 3-oxoacyl-[acyl-carrier-protein] synthase, KASII (EC 2.3.1.41) (see #7 above)

12) 54bp later, Para-aminobenzoate synthase, aminase component (EC 2.6.1.85) (contains a Mg^{2+} [Parsons JF, et al., 2002])
13) -7bp (overlap) later, Aminodeoxychorismate lyase (EC 4.1.3.38) (no known metal ion required)
14) -3bp (overlap) later, FIG004453: protein YceG like aminodeoxychorismate lyase
15) 28bp later, Thymidylate kinase (EC 2.7.4.9) ("absolute requirement for divalent cation. When Mg^{2+} is equal to ATP, the rate of dTMP kinase reaction is maximal", additionally Mn^{2+} and Co^{2+} may substitute for Mg^{2+}, with 41% and 18% of the activity acheived with Mg^{2+}, respectively [Nelson DJ, et al., 1969])
16) 86bp later, putative DNA polymerase III delta' subunit (EC 2.7.7.7) (Mg^{2+} required [Tuske S, et al., 2000]).
17) 30bp later, Putative deoxyribonuclease YcfH
18) -3bp (overlap) later, Lipid A export ATP-binding/permease protein MsbA
19) 132bp later, FUPA25 P-type ATPase
20) 121bp later, tRNA-His (convergently transcribed to the FUPA25 gene, as well as the next 2 tRNA genes)
21) 31bp further, tRNA-Arg
22) 52bp furhter, tRNA-Pro
23) 78bp further, methylenetetrahydrofolate dehydrogenase (NADP+) (EC 1.5.1.5) (inorganic phosphate required for activity, competitive inhibitor of NADP, Mg^{2+} required for activity, Mn^{2+} can substitute [Christensen KE, et al., 2008])/Methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9) (no known metal

ion required)

The next gene in the same direction of the FUPA25 gene is 1116bp away and thus is predicted to be separate from any operon containing the FUPA25 gene. Also the gene encoding Maf/YceF/YhdE family protein is transcribed in the opposite direction of the rest of the series, with the genes flanking it being 720bp apart. The genes preceding it in the list above are included because they are essentially always found with the rest of the cluster. Transcribed divergently from the first gene in this cluster, Ribosomal large subunit pseudouridine synthase C, 668bp away, is a gene encoding a ribonuclease E. Downstream of this gene is a series of genes all convergently transcribed to it, the nearest one being 98bp away, thus it is predicted to be transcribed alone as an operon, although very likely along with Ribosomal large subunit pseudouridine synthase C, as they are essentially always found transcribed divergently from one another.

RegPredict does not include *T.crunogena* at this time.

While the protein **Dvu3** (*Desulfovibrio vulgaris* str. Hildenborough (δ) 633aa, gi:46581204) was described in Chan, et al. (2010) as a FUPA25, it clusters in TCBLAST with P-type ATPase Family 6 (Heavy Metal Transporters) with an e value=e-96, then with FUPA32 with e-88 and finally with FUPA25 with e-83. Nevertheless, it is described and considered here. It is found to frequently cluster near genes encoding a transcriptional regulator of the MarR family (8bp immediately downstream here) and the thiamin biosynthesis lipoprotein ApbE (next downstream, convergent). While the *marR*

gene is the only one transcribed codirectionally to the FUPA25 gene, an operon, which may include as many as 8 genes is transcribed divergently from the FUPA25 gene, 181bp away. This putative operon encodes the following proteins (in order of transcription).

1) a 444aa hypothetical protein with 1 TMS and significant homology to PAS fold-3 domain protein [*D.vulgaris* RCH1], length=421, Expect=0.0, Identities=420/421 (99%), Positives=421/421 (100%), Gaps=0/421 (0%)
2) 248bp later, a transcriptional regulator of the GntR family
3) 280bp later, a 469aa protein with 11 TMSs and significant homology to an amino acid transporter of the AAT family [*D.alaskensis* G20], length=498, Score=944 bits (2440), Expect=0.0, Identities=430/469 (92%), Positives=457/469 (97%), Gaps=0/469 (0%)
4) 17bp later, Metal-dependent hydrolases of the beta-lactamase superfamily II
5) 842bp later in *D.vulgaris subsps vulgaris*, but only 50-80bp in other similar organisms with this gene series, a cobalamin synthesis protein/P47K family protein is encoded. The extra intergenic space in this organism replaces coding sequence in other organisms so that the protein is only 384aa long here, whereas in most similar organisms it is ~600aa long.
6) -13bp later (overlap), MotA/TolQ/ExbB proton channel family protein
7) 20bp later, biopolymer transport protein, ExbD/TolR family
8) -3bp later (overlap), TonB domain protein

RegPredict suggests several other genes and operons that may be coexpressed with those described above, as well as consistently reaffirming that the FUPA25 gene and the 444aa hypothetical protein gene are coexpressed.

The highest scoring suggested regulatory sequence preceding the FUPA25 gene is agGGggcgGCtgGCtctgCCtg (score=7.06). It is found related to a sequence (agGcaGagccagccgcCccCtg, score=5.81) preceding the aforementioned gene for a 444aa hypothetical protein, and these two are both related to 3 other sequences. The first of these is (aAGGggcgGatgcCacggCCTg, score=5.60) preceding a gene encoding a Fe-S cluster-binding protein. The next precedes a 5-gene operon encoding the following (in order of transcription):

1) membrane fusion protein of RND family multidrug efflux transporter, acrA
2) 16bp later, multidrug efflux transporter acrD/MexF
3) -3bp later (overlap), RND efflux transporter, outer membrane factor (OMF) lipoprotein, NodT family, suggested in SEED to be part of a tripartite multidrug resistance system
4) 30bp later, transciptional regulator, MarR family
5) 89bp later, a 302aa hypothetical protein with 1 TMS near the N-terminus and some homology to long-chain fatty acid transporter, putative [*Pseudomonas stutzeri* DSM 4166], length=433, Score=57.8 bits (138), Expect=9e-07, Identities=45/210 (21%), Positives=71/210 (34%), Gaps=35/210 (17%)

The last related sequence in this group is agGctGcgGttggCgcCctCtg (score=5.00) and it precedes a gene encoding a GCN5-related N-acetyltransferase, GNAT family.

The next suggested regulatory sequence preceding the FUPA25 gene is aaaaAtatCaGgagTcggc (score=6.67). It is found related to a sequence (AaaGATcacatcaATCcgT, score=5.32) preceding the aforementioned gene for a 444aa hypothetical protein, and these two are both related to another sequence (aCacatcaCaGgaggcgGc, score=4.99) preceding a bicistron encoding a nicotinamidase/isochorismatase family protein, with specific hits in the Cysteine hydrolases conserved domain, followed by a DSBA-like thioredoxin domain protein, possibly an oxidoreductase.

Another suggested regulatory sequence preceding the FUPA25 gene is aGggGgcggctggctCtgCc (score=6.67). It is found related to a sequence (aGGcagaGCcaGCcgccCCc, score=5.36) preceding the aforementioned gene for a 444aa hypothetical protein, and these two are both related to another sequence (GGggGgcGgcagCcgCggCC, score=5.36) preceding a 7-gene encoding the following proteins (order of transcription):

1) glycolate oxidase 2-subunit type (EC 1.1.99.14) (no known metal cofactors), Fe-S subunit GlcD
2) 32bp later, glycolate oxidase 2-subunit type (EC 1.1.99.14), Fe-S subunit GlcF

3) 124bp later, phosphate acetyltransferase (EC 2.3.1.8) (enzyme is activated by low concentrations of univalent cations, sequence of effectiveness: NH_4^+, K^+, Na^+ [Kyrtopoulos SA., 1973])
4) 26bp later, acetate kinase
5) 100bp later, a 363aa hypothetical DRTGG domain protein with no TMSs and homology to cobyrinic acid a,c-diamide synthase family protein [*D.vulgaris str. Hildenborough*], length=354, Score=328 bits (841), Expect=1e-108, Identities=164/347 (47%), Positives=243/347 (70%), Gaps=8/347 (2%)
6) -3bp later (overlap), a 209aa hypothetical protein (predicted to be L-lactate dehydrogenase subunit YkgG by RegPredict) with no TMSs and with homology to Lactate utilization protein B/C [*D.alaskensis* G20] length=205, Score=290 bits (743), Expect=2e-98
   Identities=140/205 (68%), Positives=168/205 (82%), Gaps=2/205 (1%)
7) 4bp later, cysteine-rich domain-containing 4Fe-4S-binding Lactate utilization protein B/C (predicted to be L-lactate dehydrogenase Fe-S oxidoreductase subunit YkgE by RegPredict)

Another related sequence (GGgcGgcGCg|tGCcgCcgCC, score=5.09) precedes a single gene encoding a arginine biosynthesis bifunctional protein ArgJ, glutamate N-acetyltransferase (EC 2.3.1.35) (no known ion requirements)/N-acetylglutamate synthase (EC 2.3.1.1) (no known ion requirements).

The next suggested regulatory sequence preceding the FUPA25 gene is AagAtGtTAtttTAtCtTgaT (score=6.65), with a related sequence (cTcCtgAtAtttTtTatGcAt, score=5.63) preceding the aforementioned gene for a 444aa hypothetical protein, and these two are both related to another sequence (AagAtGgtgttttctCtTtaT, score=4.99) preceding a 7-gene operon encoding the following proteins (order of transcription):

1) Chemotaxis response regulator containing a CheY-like receiver domain and a protein-glutamate methylesterase domain, CheB-2 (EC 3.1.1.61) (phosphorylation of intact enzyme requires Mg^{2+} [Anand GS, et al., 1998]).
2) 59bp later, HEAT repeat-containing PBS lyase
3) -3bp later (overlap), chemotaxis protein methyltransferase, cheR-2 (EC 2.1.1.80) (Ca^{2+}, Mn^{2+}, Cu^{2+}, Zn^{2+}, Co^{2+}, Fe^{2+}, Fe^{3+} and Mg^{2+} at 2 mM have no effect on enzyme activity of methyltransferase I, but Mg^{2+} and Ca^{2+} activate MCP methyltransferase II [Kim S, 1984])
4) 51bp later, Cobyrinic acid a,c-diamide synthase, ParA family protein
5) 90bp later, positive regulator of CheA protein activity, chemotaxis protein CheW
6) 10bp later, Chemotaxis response regulator receiver protein - transmits chemoreceptor signals to flagellar motor components, CheY-2
7) 96bp later, signal transduction histidine kinase chemotaxis protein, CheA (EC 2.7.13.3) (Heme/Iron, Mg^{2+} cofactors [Shrivastava R, et al., 2007], Mg^{2+} also may be required [Del Papa MF, et al., 2008])

The next suggested regulatory sequence preceding the FUPA25 gene is CttgAtTgCaaaGtAtTtttG (score=6.65), with a related sequence (CtttGatacaaaaataCtttG, score=5.24) preceding the aforementioned gene for a 444aa hypothetical protein, and these two are both related to another sequence (attgctTgcaaaatAttgttg, score=4.98 preceding a single gene encoding a probable poly(beta-D-mannuronate)/ alginate O-acetyltransferase, AlgI.

The next suggested regulatory sequence preceding the FUPA25 gene is AgaTTTcATGCATaAAAaaT (score=6.54), with a related sequence (AttTTTtATGCATgAAAtcT, score=6.02) preceding the aforementioned gene for a 444aa hypothetical protein, and these two are both related to another sequence (ATTTTTcacgaaaaAAAAAT, score=5.05) preceding a 8-gene operon encoding the following proteins (order of transcription):

1) a 578aa hypothetical protein with no significant homology or TMSs, particular to the Desulfo- clades of δ-proteobacteria
2) 39bp later, Aminomethyltransferase (glycine cleavage system T protein), gcvT (EC 2.1.2.10) (no known cation requirements)
3) 52bp later, 16S ribosomal RNA methyltransferase, RsmE
4) -3bp later (overlap), AAA ATPase central domain-containing recombination factor protein, RarA
5) -3bp later (overlap), glycosyl transferase, group 2 family protein

6) -3bp later (overlap), 1-acyl-sn-glycerol-3-phosphate O-acyltransferase (EC 2.3.1.51) (appears to lack Mg^{2+} requirement found in eukaryotes [Okuyama H, et al., 1973])
7) 185bp later, a 61aa hypothetical protein found only in this organism (no homologues in NCBI) and no longer a valid gene model in the MicrobesOnline database
8) 198bp later, transcriptional regulator, TetR family*

   *because #7 is no longer considered valid in MO, this gene, while predicted to be part of the operon by RegPredict, is no longer considered close enough to upstream coding sequence to be part of the same operon by MO.

Firmicutes:

TC 3.A.3.25.3

Efa12 (*Enterococcus faecalis* V583 601aa, gi:29377108) is encoded by a FUPA25 gene located approximately half-way along an expansive stretch of sequence shared with *Bacteroides vulgatus* EK4 (84.8kb in *E.faecalis*, 82.7kb in *B.vulgatus*). With the exception of an occasional hypothetical protein gene and one mobile element, this region encodes all the same genes in both organisms. The gene is predicted by MicrobesOnline to be expressed alone, though it is sandwiched amongst an extensive swath of codirectionally transcribed genes. As such, several clustering genes upstream and downstream are noted here for convenience. They encode the following, in order of transcription:

1) Hypothetical protein (identical to N-terminus of larger PhoP (below) in most homologues, probably a protein fission or sequencing error)
2) 25bp later, Alkaline phosphatase synthesis transcriptional regulatory protein PhoP
3) 24bp later, hypothetical protein NCBI BLAST hits ORF 73; extensive acidic domains, potential leucine zipper; immediate early protein homologue
4) 17bp later, N-acetylmuramoyl-L-alanine amidase, family 4
5) 220bp later, cell wall teichoic acid glycosylation protein GtcA
6) 72bp later, Nicotinate phosphoribosyltransferase (EC 2.4.2.11)
7) 3bp later, NAD synthetase (EC 6.3.1.5)
8) 192bp later, FUPA25 P-type ATPase
9) 92bp later, 95aa hypothetical protein with no clear TMSs and homology to YjdI-like protein [*Lactococcus lactis subsp. lactis*], length=90, Score=122 bits (306), Expect=2e-35, Identities=55/90 (61%), Positives=66/90 (73%), Gaps=0/90 (0%)
10) -10bp later (overlap), acetyltransferase, GNAT family
11) 168bp later, Preprotein translocase subunit SecG (TC 3.A.5.1.1)

The 192bp upstream of the FUPA25 gene in *E.faecalis* and *B.vulgatus* were used to search for related sequences in the *E.faecalis* genome. Although the most recently updated version of this genome in SEED, as well as that in RegPredict and MicrobesOnline, suggest that a 31aa peptide is encoded 26bp upstream of this FUPA25 gene, the complete lack of homology to this suggested coding region, both within SEED

as well as in NCBI BLAST, gives cause to disregard this proposed peptide at this point in time.

RegPredict suggests several palindromic sequences found upstream of the FUPA25 gene in this organism as possible regulatory sequences. Four such sequences are discussed here.

The first such suggested FUPA25 regulatory sequence (tCtTTTTCaaaacaGAAAAgGt, score=8.12) is found 63bp upstream of the FUPA25 gene. Two sequences related to this are found, the first of which (TatTTTTaaaaaaagAAAAggA, score=6.15, 32bp upstream) precedes a bicistron encoding triosephosphate isomerase (EC 5.3.1.1) (no metal cations are known to be required [Mathur D, et al., 2006]) and enolase/phosphopyruvate hydratase (EC 4.2.1.11) (two Mg^{2+} per subunit are required for catalytic activity [Hosaka T, et al., 2003]). The second sequence in this series (tttTTTTcAcaaaTaAAAAggt, score=5.66) precedes a tricistron encoding three components of an D-methionine/metal ion import ABC transporter: the ATP-binding protein metN, the permease protein and the substrate binding protein, D-methionine-binding lipoprotein metQ.

The second such suggested FUPA25 regulatory sequence (GgTcTGTgttttttACAtAaC, score=8.12) is found 163bp upstream of the FUPA25 gene. The related gene sequence (GgTcTGtgCttttGatCAaAaC, score=5.66) precedes a gene encoding Xanthine/uracil/thiamine/vitamin C permease.

The third suggested FUPA25 regulatory sequence (cacagacCatCGgcGaatgaac, score=8.12) is found 188bp upstream of the FUPA25 gene. The related sequence

(CactgaaCgtcagtGaatgaaG, score=5.17) precedes a bicistron encoding a membrane-associated zinc RIP metalloprotease RseP followed by prolyl-tRNA synthetase (EC 6.1.1.15) (requires Mg^{2+} [Crepin T, et al., 2006]). Another related sequence (caAAgaccatCGgacaaaTTac, score=5.17) precedes another biscistron, this one encoding a cell division transporter, ATP-binding protein FtsE (TC 3.A.5.1.1) and cell division protein FtsX.

The fourth suggested FUPA25 regulatory sequence (AtgAACgatGgtCtgtGTTttT, score=8.12) is found 172bp upstream of the FUPA25 gene. Four related sequences are found and described here. The first related sequence (AtcatCgcttgtctttGtgttT, score=5.17) precedes a bicistron encoding another P-type ATPase, of Family 6 (Heavy Metal transporters), with 634aa and 7 TMSs, followed 35bp later by Mg^{2+} transporter-C (MgtC) family protein SapB. Another possibly related sequence (AtAAAggaAtATcTgttTTTtT, score=5.17) precedes a 4-gene operon encoding the following (in order of transcription):

1) outer surface protein of unknown function, cellobiose operon
2) N-acetylmuramic acid 6-phosphate etherase (EC 4.2.-.-)
3) PTS system, IIBC components (EC 2.7.1.69) (no known cation requirements)
4) phosphosugar-binding transcriptional regulator, RpiR family

The third sequence is in fact two sequences both found upstream of the very same Xanthine/uracil/thiamine/vitamin C permease described above. The sequences, ctgtGCtttGgtCtgtGCtttt and attAaaAaAGgtCTgTgcTttg (both scores=5.17), are found 195 and 206 bp upstream of the gene, respectively.

The last sequence suggested to be related (aggaAAAatGgtCtgTTTatta, score=5.17) precedes a single gene encoding an ATP-dependent Clp protease, ATP-binding subunit ClpE.

The gene encoding **Ljo12** (*Lactobacillus johnsonii* NCC 533 623aa, gi:42519841) is found between two other codirectionally transcribed genes, encoding a cystathionine beta-synthase (CBS) domain protein 70bp upstream and a Na^+:H^+ antiporter NapA 172bp downstream. The CBS domain protein gene is found in about the same proximity to the FUPA25 gene in numerous Lactobacillus species. Upstream of the CBS domain protein gene is a 326bp gap before the next gene, encoding a 94aa hypothetical protein no significant homology, which greatly reduces the likelihood that the predicted operon begins before the CBS domain protein gene. Downstream of the NapA encoding gene there is a gene only 37bp away, but it is transcribed convergently. This gene encodes a 217aa hypothetical protein with 1 TMS near the N-terminus. Perhaps more useful and relevant is a gene directly upstream of this, which is 0bp away, and thus cotranscribed, encoding an ABC transporter ATPase and permease components. SEED finds homologues of this gene to be the most frequent to cluster near homologues of the gene encoding Ljo12, so while the genes do not share an operon, it is likely that their colocalization implies coexpression.

RegPredict suggests several palindromic sequences found upstream of the FUPA25 gene in this organism as possible regulatory sequences. Five such sequences are discussed here.

The first such sequence (AtattttGaAaTcCgcgtctT, score=6.66) preceding the FUPA25 gene has one related sequence (AActTttGaAaTcCggAttTT, score=5.23) precedes a 16-gene operon which encodes the following in order of transcription:

1) a 111aa hypothetical protein with 4 TMSs with homology to lipoprotein, putative [*Carnobacterium sp*. AT7] length=124, Score=53.7 bits (128), Expect=9e-08, Identities=28/111 (25%), Positives=57/111 (51%), Gaps=3/111 (3%)
2) 180bp later, cell division protein MraZ
3) -16bp later (overlap), rRNA small subunit methyltransferase H/16S rRNA m(4)C1402 methyltranserfase/S-adenosyl-methyltransferase MraW
4) 29bp later, cell division protein FtsL
5) 0bp later, cell division protein FtsI/penicillin-binding protein 2B (EC 2.4.1.129) (slight stimulation with Mg^{2+}, but no divalent cation required for function [Nakagawa J, et al., 1984])
6) 25bp later, phospho-N-acetylmuramoyl-pentapeptide-transferase MraY (EC 2.7.8.13) (Mg^{2+} required, can be substituted, albeit poorly, by Mn^{2+} [Bouhss A, et al., 2004])
7) 8bp later, UDP-N-acetylmuramoylalanine--D-glutamate ligase (EC 6.3.2.9) (highly dependent on Mg^{2+} and K_3PO_4 [Pratviel-Sosa F, et al., 1991])
8) 3bp later, UDP-N-acetylglucosamine--N-acetylmuramyl- (pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase (EC 2.4.1.227)
9) 19bp later, Div1B-like protein, Cell division protein FtsQ
10) 65bp later, cell division protein FtsA

11) 18bp later, cell division protein FtsZ
12) 19bp later, cell division protein sepF/FtsZ-interacting protein related to cell division
13) 0bp later, cell division protein YlmG/Ycd19 (putative), YggT family
14) 2bp later, ribosomal S4e-like ribosomal binding domain cell division protein
15) 5bp later, cell division initiation protein DivIVA

The second such sequence (acAaaGtAaTaAaTtCaaTta, score=6.62) preceding the FUPA25 gene has one related sequence (TtAaagacaTaAatataaTtA, score=5.09) precedes a single gene that encodes a M42 family glutamyl aminopeptidase I zinc metalloprotease.

The third such sequence (cGcTaTTAgcatTAAcAtCc, score=6.18) preceding the FUPA25 gene has one related sequence (TacTaTtAtcatTtAcAtaA, score=5.31) preceding a bicistron encoding an ABC transporter ATPase component and an ABC transporter antimicrobial peptide permease protein.

The fourth such sequence (CTtTcATata|atgATaAtAG, score=5.88) preceding the FUPA25 gene has five related sequences, The **first** of which (CattTAcATaaATtTAggaG, score=4.92) precedes the same bicistron encoding the ABC transporter components just previously described above.

The **second** related sequence in this set (gTttTttAtaatTttAgtAa, score=5.16) precedes a 4-gene operon encoding a PTS system cellobiose-specific IIC component (EC 2.7.1.69), a phosphatidylserine decarboxylase (EC 4.1.1.65) (stimulated by the divalent

cations Mn^{2+}, Mg^{2+}, Ca^{2+} and Ba^{2+}, in order of increasing stimulation [Verma JN, et al., 1985]), an arginine/ornithine antiporter ArcD and another phosphatidylserine decarboxylase of the exact same length as that encoded two gene before it and 87% identical. The **third** related sequence (CTtTTAAataataTTAAtAG, score=5.10) precedes a 3-gene operon encoding a heptaprenyl diphosphate synthase component II (EC 2.5.1.30) (Mg^{2+} required, Mn^{2+} also works but with only 30% of the activation obtained with Mg^{2+} [Suzuki T, et al., 2006]), a major facilitator superfamily permease and a 1-deoxy-D-xylulose-5-phosphate synthase (EC 2.2.1.7) (divalent cation Ligand req'd; Mg^{2+}, Mn^{2+} work best, Zn^{2+} to a lesser extent [Bailey AM, et al., 2002]). The **fourth** related sequence in this set (tTtTtATaaaaagATgAtAg, score=5.08) precedes a single gene encoding a levansucrase precursor with homology to fructosyltransferase Ftf [*L.reuteri*], Length=798, Score=873 bits (2255), Expect=0.0, Identities=448/818 (55%), Positives=578/818 (71%), Gaps=42/818 (5%). The **fifth** and last related sequence (CtttttTAtaatTAtttttG, score=4.67), though of relatively low score, is of interest due the recurrence of the genes it precedes in searching through numerous sequences related to those preceding the FUPA25 gene. The genes comprise a 27-gene operon encoding various Lj928 prophage proteins including small and large terminase subunits, portal proteins, major and minor tail proteins, holin and lysin. While these genes may not directly suggest what the cell uses its FUPA25 protein for, the possibility that Lj928 prophage may use similar regulatory sequences, and thus a similar set of transcription factors, may assist in discovering the function(s) of the FUPA25 protein indirectly.

The fifth and final such sequence (GCgCggAttttttTgtGgGC, score=6.17) preceding the FUPA25 gene has one related sequence (aCgCggAtTttAtTgaGgGg, score=5.00) preceding a 24-gene operon encoding 16 ribosomal subunit (the 30S protein S14p/S29e is of note due to its dependency on zinc [van der Kaaij H, et al., 2004]), a preprotein translocase secY subunit (TC 3.A.5.1.1), adenylate kinase (EC 2.7.4.3) (divalent cations required to form a complex with di- or trinucleotide; in decreasing order of efficiency: Mg^{2+}, Ca^{2+}, Mn^{2+}, Ba^{2+}; enzymatic reaction resembles inorganic metal catalysis [Noda L, 1973]), translation initiation factor 1, DNA-directed RNA polymerase alpha subunit (EC 2.7.7.6), ATPase and transmembrane components of the general energizing module of ECF transporters and a tRNA pseudoserine synthase A (EC 5.4.99.12) (no known metal cation requirements in prokaryotes).

There is a notable tendency for genes encoding the FUPA25 proteins to be associated either by colocalization or predicted co-regulation with DNA-directed RNA polymerase subunits, DNA polymerase subunits, cell division proteins, ABC-type transporter subunits and ribosomal subunits. This data, taken collectively, suggests that the FUPA25 proteins are involved in helping to regulate and/or supply the mechanisms of cell division or perhaps protein expression by specifically supplying them with necessary divalent cations. Both of these processes are also likely candidates for the prophage described above to piggyback onto their regulation.

The FUPA26 ATPase Family

FUPA26 is a family of Type I P-type ATPases with 3 members described in Chan, et al. (2010) of sizes ranging from 841-900 aas in length and a topology of 8 TMSs with 1 & 2 pairing and found around 180 aas, 3 & 4 nearby, pairing around 244 aas, 5 & 6 pairing around 445 aas and 7 & 8 pairing around 795 aas. The family's nearest hit in TCDB is Family 5 (3.A.3.5.-), then Family 6 (3.A.3.6.-), but both are considerably more distantly related than other FUPAs are from their respective hits. For example, a TCBLAST of a FUPA25 gets its best outside-family hit with Family 6 with an e-value of e-72, whereas a TCBLAST of a FUPA26 gets its best outside-family hit with Family 5 with an e-value of e-51.

Additionally, SEED found 27 other FUPA26 members, all in genus Corynebacterium, although 14 are in duplicate genomes. Thus, there are 13 other unique members, as well as the original 3, all found in very similar gene neighborhoods, and as such, will mostly be described together, mentioning exceptions as necessary. The expanded set of FUPA26s range in size from 674 aas (in *C.genitalium*) to 948 aas (in *C. nuruki*), as well as an incomplete fragment missing the first 300 residues, and consequently TMSs 1-4, in *C.lipophiloflavum*. It may be also be of some interest to find that 7 members of this genus contain no FUPA26, specifically *C.pseudogenitalium, C.aurimucosum, C.striatum, C.tuberculostearicum, C.accolens* and *C.ammoniagenes*. Whether this is an indication that FUPA26 is subjectively non-essential or supplemented in some way by other transporters has yet to be determined.

Cdi2 (*C.diphtheriae* NCTC 13129 841aa, gi:38233019), **Cef6** (*C.efficiens* YS-314 976aa, gi:25027013) and **Cgl2** (*C.glutamicum* ATCC 13032 878aa, gi:19551679) were originally used as the basis for describing the FUPA26 family in Chan, et al. (2010). It should be noted that **Cef6** is 900aa long, despite the notation in Chan, et al. (2010). The FUPA26 genes encoding these three proteins, as well as those encoding all other FUPA26 proteins known, are nestled in well-conserved gene neighborhoods. The FUPA26 gene is typically the sixth to ninth gene in a long operon in all instances, with another 7-9 genes following it, although genes transcribed convergently to the operon sometimes separate these. The most typical arrangement encodes as follows, in order of transcription:

1) Glutaredoxin-like domain protein
2) ~230bp later, Glutamyl-tRNA reductase (EC 1.2.1.70) (Mg^{2+} stimulates activity and restores it after treatment with chelating agents, Ca^{2+} and Mn^{2+} also restore activity [Schauer S, et al., 2002])
3) -3 to 3bp later, Porphobilinogen deaminase (EC 2.5.1.61) (no known metal cofactor requirements)
4) 70 to 255bp later, Uroporphyrinogen-III methyltransferase (EC 2.1.1.107) (no known cation requirement)/ Uroporphyrinogen-III synthase (EC 4.2.1.75) (no ion requirement observed [Alwan AF, et al., 1989])
5) 30 to 60bp later, Porphobilinogen synthase (EC 4.2.1.24) (Mg^{2+} may be required [Yamasaki H, et al., 1971])

6) 5 to 35bp later, 3/4-TMS conserved FUPA26-associated hypothetical protein
7) 6 to 245bp later, TerC family integral membrane protein
8) -7 to 245bp later, FUPA26
9) 11 to 268bp later, Uroporphyrinogen III decarboxylase (EC 4.1.1.37) (possible slight activation with Mn^{2+} [Jones RM, et al., 1993]).
10) 1 to 68bp later, Protoporphyrinogen IX oxidase, aerobic, HemY aka HemG (EC 1.3.3.4) (no known cation requirements).
11) 50 to 108bp later, Glutamate-1-semialdehyde aminotransferase (EC 5.4.3.8) (no metal ion requirement observed [Mayer SM, et al., 1994])
12) -3 to 141bp later, phosphoglycerate mutase/fructose-2,6-bisphosphatase (EC 5.4.2.1) (requires Mn^{2+} for activation [Kuhn NJ, et al., 1993])
13) 1 to 60bp later, Thiol:disulfide oxidoreductase related to ResA
14) -7 to 55bp later, Cytochrome c-type biogenesis protein CcdA (DsbD analog)
15) 1 to 162bp later, Ccs1/ResB-related putative cytochrome C-type biogenesis protein
16) 9 to 175bp later, Cytochrome c-type biogenesis protein CcsA/ResC
17) 0 to 96bp later, dTDP-glucose 4,6-dehydratase (EC 4.2.1.46) (Mg^{2+} may be required [Singh B, et al., 2010])

While other genes do occasionally occur after or with this arrangement, there is little pattern or significant frequency of this. The two most frequently occurring, sometimes together, are operons encoding proteins in the ABC-transporter petrobactin-

mediated iron uptake system (all have top hits in TCBLAST in TC# 3.A.1.14.-) and those in the dTDP-rhamnose synthesis system. There are also two frequently occurring divergently transcribed operons that likely share a regulatory region with the FUPA26-containing operons. The most frequently occurring encodes a single protein, phosphoserine phosphatase (EC 3.1.3.3) (Mg^{2+} binding, Mn^{2+} and Co^{2+} can substitute [Schmidt LS, et al., 1973]) and is found in every Corynebacterium species except *C.glutamicum, C. efficiens, C.kroppenstedtii, C nuruki* and *C.variabile*. The former two of these species have instead 3- and 6-gene operons respectively. Both contain genes encoding Fe^{3+}/thiamine transport system, secreted component; ABC transporter substrate-binding protein followed 22-77bp later by a 12 TMS ABC-type transporter permease component with hits in TCBLAST of e-17 in TC# 3.A.1.10.1, Iron(III)-transport system permease protein sfuB and

NCBI BLAST hits for probable ferric iron transport system permease protein [*Roseovarius sp.* HTCC2601] Length=568, Score= 278 bits (711), Expect=5e-84, Identities=184/504 (37%), Positives=282/504 (56%), Gaps=22/504 (4%) and Iron(III) ABC transporter, permease protein [*Polymorphum gilvum* SL003B-26A1] Length=569, Score= 268 bits (685), Expect=3e-80, Identities=183/538 (34%), Positives=276/538 (51%), Gaps=35/538 (7%), followed 1-6bp later by putative ABC transporter ATP-binding protein with hits in TCBLAST of e-49 (highest) in TC#3.A.1.6.3, Sulfate/thiosulfate import ATP-binding protein cysA but also e-38 in TC#3.A.1.10.3, Ferric cations import ATP-binding protein fbpC2 and NCBI BLAST hits for sulfate ABC transporter, ATP-binding protein (highest) [*C.ammoniagenes* DSM 20306]

Length=350, Score= 369 bits (948), Expect=5e-125, Identities=203/352 (58%), Positives=250/352 (71%), Gaps=5/352 (1%), but also iron (Fe^{3+}) ABC superfamily ATP binding cassette transporter, ABC protein [*Desmospora sp*. 8437] Length=370, Score=231 bits (590), Expect=9e-71, Identities=135/322 (42%), Positives=195/322 (61%), Gaps=6/322 (2%).

Multiple hits against spermidine/putrescine import ABC transporter ATP-binding protein PotA also occur between these two scores, as well as numerous hits for ABC transporter ATP-binding proteins with unidentified substrates, however this may not be particularly important, as it is the ATP-binding part of the complex. In addition to these three genes in *C. efficiens*, one encoding putative Abortive infection protein is found 109bp upstream and two encoding putative L7/L12 family ribosomal protein and IS4 family mobile element protein, are found 17bp and then 73bp downstream, respectively.

As for the latter 3 exceptions, *C.kroppenstedtii* simply lacks the phosphoserine phosphatase gene and has an exceptionally large space between the Glutaredoxin-like domain protein gene and the Glutamyl-tRNA reductase gene (541bp), making it considerably less likely to be in the same operon. A similar situation is the case for *C nuruki* and *C.variabile*, but the large intergenic space (462bp and 665bp, respectively) precedes the Glutaredoxin-like domain protein gene instead.

SEED also provides evidence that similar operons to the one described above are found surrounding Family 5 (typically Cu^{+}) P-type ATPase genes, in approximately the same location as the FUPA26 genes, in *Sanguibacter keddieii, Xylanimonas cellulosilytica, Isoptericola variabilis, Cellulomonas fimi, Cellvibrio gilvus,*

Beutenbergia cavernae and *Jonesia denitrificans*, all members of Suborder Micrococcineae. Even more informative is the finding that the genes encoding Glutamyl-tRNA reductase, Porphobilinogen deaminase, Uroporphyrinogen-III methyltransferase / Uroporphyrinogen-III synthase and Porphobilinogen synthase, mostly involved in heme/siroheme biosynthesis, are found just preceding a Family 5 gene in *Frankia sp.* Ccl3 and *F.alni* ACN14a. These Family 5 members' closest hits in TCBLAST are in 3.A.3.5.18 and 3.A.3.5.19. While the former has only been hypothesized to be a Cu^{+} transporter, the evidence is by no means conclusive, especially as the protein being referred to as CopA in one study (Villafane AA, et al., 2009) actually refers to an 80aa protein, not the ~800aa, full Family 5 ATPase.. Moreover, the latter has been explicitly characterized as having a capacity to transport Fe^{3+} (Sitthisak S, et al., 2007). Replacing the genes encoding 3/4-TMS conserved FUPA26-associated hypothetical protein and TerC family integral membrane protein is a gene encoding a 75aa protein referred to in SEED as putative Copper Chaperone but in NCBI as heavy metal transport/detoxification protein. It has strong homology to residues 354-430 in 3.A.3.5.3, 2-71 in 3.A.3.5.18 and 5-65 in 3.A.3.5.5. These gene-neighborhood and operon homologies strongly suggest one of two possibilities; that the FUPA26 family is a specialized supgroup of Cu^{+} transporters, possibly functioning simply to rid the cell of excess copper, or that FUPA26, as well as the aforementioned members of Family 5, are Fe^{3+} transporters for delivering the cation to the heme/siroheme biosynthesis pathway for subsequent incorporation into cytochrome c. This specialization may have originated from a copper/iron toxicity response pathway, thus explaining the sensitivity of

3.A.3.5.18 to Cu^{+}, or it may be that this pathway is sensitive to some degree to both Fe^{3+} and Cu^{+}, as both are necessary for electron transfer through the interaction of cytochrome c [Fe^{3+}] and cytochrome c oxidase [$2Cu^{+}$].

RegPredict suggests a number of accessory proteins, mostly involved in ferric iron transport, heme/siroheme and cytochrome c biogenesis, to be regulated by sequences related to those preceding the FUPA26-containing operons. Several examples follow below.

Related sequences such gAtGaggtggaacCgTg (score=4.14), GaagtggaccagctggC (score=5.53) and GatgcgctccagctggC (score=4.81) are exemplary of one set found preceding the operons containing the FUPA26 genes. Several other genes of interest are found in conjunction with these genes, either under the regulation of related sequences or co-localized with genes commonly found in the general FUPA26 operon when those genes are separated from the operon. For example, in *C.aurimucosum*, a second related sequence is found (GatTCggCgGaaGAggC, score=4.32), this one preceding a gene encoding hemoprotein HemQ which is a chlorite dismutase which is an essential component of the heme biosynthetic pathway in Gram-positive bacteria, far removed from the FUPA26 encoding operon. A similar sequence precedes the same gene in *C.amycolatum* (GagttggacgaggtgcC, score=4.10) and in *C.kroppenstedtii* (gaggcggcgaagaaggt, score=3.79), where it is preceded by genes encoding Uroporphyrinogen III decarboxylase (EC 4.1.1.37) (possible slight activation with Mn^{2+} [Jones RM, et al., 1993]) and Protoporphyrinogen IX oxidase, aerobic, HemY aka HemG (EC 1.3.3.4) (no known cation requirements), which are missing from the organisms'

FUPA26-containing operon. In *C.jeikeium* there is a similar case, as the same genes are translocated, but here the sequence (GagGccGaccCaaCggC, score=4.04) precedes the Protoporphyrinogen IX oxidase, *C.nuruki* and *C.variabile* may well have similar arrangements, as the same genes are absent from the FUPA26-containing operon, but these genomes are currently not included in RegPredict. The genes encoding the Petrobactin ABC transporter components are also sometimes separated from the FUPA26-containing operon, but a related sequence precedes them where they are found elsewhere in the genome. This can be seen for this sequence in *C.urealyticum*, in which two related sequences (GagctggTcAagcagaC, score=3.74 and gatgCggccaagGagga, score=4.22) precede a version of 3 of the aforementioned genes encoding permeases I and II and the second ATP-binding (NBD)/ATPase component. Similarly, in *C.aurimucosum*, a minor related sequence (gaagCgGcgtCgGaggg, score=3.72) is found preceding a six-gene operon encoding a putative aminobenzoyl-glutamate transporter and a putative peptidase, followed by 4 gene-encoding the 2 pairs of Petrobactin ABC transporter components.

Another set of predicted regulatory sequences suggested by RegPredict precede the gene encoding uroporhyrinogen-III methyltransferase/synthase in the FUPA26 gene encoding oepron. The sequences, in *C.glutamicum* (attgTGcgCcGacCAgtta, score=4.02 and gctGTtgatggagtACgtt, score=3.58), in *C.diptheriae* (AtGgTGcTgcgAcCAtCtT, score=3.97), in *C.jeikeium* (actgtGCatctaGCcgcta, score=5.10) and in *C.kroppenstedtii* (gtttTGgattgagCAcgta, score=4.21) also have related sequences of interest preceding likely coregulated genes elsewhere in the respective genomes (it is worth noting that the

Corynebacterium genomes not mentioned above also had related sequences at the same locus. However, the scores for these were below the preferred threshold and thus not used as the standard for comparisons of potential relatedness). Related sequences preceding 3-4 gene operons encoding ABC-type transporters are found in *C.glutamicum* (cAttTcaatctaaccAgtTa, score=3.59), in *C.aurimucosum* (AcggtaaggttacccgttT, score=3.53), which lacks a FUPA26, but not the rest of the operon in which it is usually found, and in *Frankia sp*. Ccl3 (ccggTcCTgctAGcAggtc, score=4.07). The proteins they encode are variably identified as Zn^{2+}, Mn^{2+} or cobalamin/Fe^{3+} sideophore transporters, however TCBLAST analysis reliably shows highest hits for the ATPase components mostly in TC # 3.A.1.15.8, 3.A.1.15.5 and several TC# 3.A.1.14.-, and for the permease components all in TC# 3.A.1.15.-, primarily 3.A.1.15.6 and 3.A.1.15.2.

Another pair of predicted signal sequences preceding the operon containing the FUPA26 gene identified by RegPredict best exemplified in *C.glutamicum* (tgttGatggagtaCgttg, score=5.67) and *C.efficiens* (gggCggtgctgtgaGttg, score=5.81) (but present with lower scores in several other Corynebacteria) is found preceding the gene encoding Uroporphyrinogen-III methyltransferase/synthase. In *C.amycolatum*, a related sequence (tgGtGgtgccgcaCgCtg, score=4.70) precedes a bicistron encoding the heme ABC import ATP-binding/ATPase component HmuV and heme oxygenase protein HmuO (EC 1.14.99.3) (Fe^{2+} required [Furci LM, et al., 2007]). HmuO is also preceded by related sequence in *C.glutamicum* (tgttgatgaagtcttctg, score=4.18) and *C.diptheriae* (GgttgatgtggtgaggtC, score=4.32), albeit with lower scores. Lastly for this related sequence set, *C.amycolatum* contains a sequence (ggttgatgctgtggcgtg, score=4.84) which

precedes a 5-gene operon encoding 4 ABC transporter components, all of which have their highest hits in TCBLAST in 3.A.1.15.-. While several of these hits are in subfamilies with uncertain Fe^{3+} transport capacity, the fifth gene encodes an iron repressor protein of the diptheria toxin repressor of the family DtxR.

A considerable wealth of data regarding the regulatory environment of the FUPA26 encoding genes and their neighbors implies that these gene encode either Fe^{3+} direct transporters, or transporters of iron-sideophore complexes. The ultimate destination of the iron appears to primarily be a heme biogenesis complex, which synthesizes heme for use in cytochrome c biogenesis. With an array of both P-type ATPases and ABC transporters expressed for this task, the cell covers both low affinity/high specificity and high affinity/low specificity ranges. This broad coverage may be necessary both as a pathogen needing to extract bound iron from its host and in a free-living soil environment in which competition for resources is fierce.

The FUPA27 ATPase Family

FUPA27 P-type ATPases are generally considered to be of type I topology. They are closest to Family 5 (Copper Transporters) of the known-function P-type ATPases. Their differences however have given rise to the categorization as type VIII ATPases, along with FUPA29 (Chan, et al., 2010, p.39), because while they are of the typical type I size, many of them appear to differ slightly in topology. They are predicted by the WHAT program (but not by the TMHMM program) to have one or two additional TMS preceding TMS A [Chan et al., 2010]. They are found in α-, β-, γ- and ε-Proteobacteria as well as in some Spirochaetes; specifically, Leptospira. The general topology of this family consists of 8, 9 or (possibly) 10 TMSs, depending on if the 2 possible TMSs sometimes preceding TMS A are absent. In any case, A & B cluster together within a 60aa region within the range of residues 152-242, TMSs 1 & 2 cluster together within a 45aa region within the range of residues 217-293, TMSs 3 & 4 cluster about 20 residues from the center of the protein, within residues 403-482 and 5 & 6 cluster within the last ~60 residues, ~754-809. The “extra” TMS preceding TMS A, referred to here as A`, can be found anywhere from 15-75aa ahead of TMS A. Where exceptions do occur, such as the unusually short FUPA27 encoded in the Sulfurimonas denitrificans DSM 1251 genome, the abrogations occur primarily in the hydrophilic region preceding TMS A` but also between it and the regular TMS A. Interestingly, when these comparisons are made using HMMGAP, it can also be noted that the predicted TMSs of the *S.denitrificans* are consistently larger than those of the FUPA27 proteins described in TCDB.

The family has 28 members represented in Chan, et al. (2010), with one per genome except for two each in the α-Proteobacteria (α-P), *Mesorhizobium loti* and *Sinorhizobium meliloti*, for a total of 26 genomes. Since FUPA27s have since been discovered in ε-Proteobacteria, *Nitratiruptor sp*. SB155-2 and *S.denitrificans* DSM 1251 will also be considered here to represent them, bringing the total to 30 representative members described here. SEED analysis was used to find hundreds of other FUPA27s encoded in other bacteria, but with very little variation from what has previously been described, the representative set need not be expanded. The FUPA27s of the α-Ps (TC# 3.A.3.27.1) range in size from 724-763aa, those of the β-Ps (TC# 3.A.3.27.1, 3.A.3.27.4) from 817-851aa, those of the γ-Ps (TC# 3.A.3.27.1) from 769-882aa, those of the ε-Ps (TC# 3.A.3.27.3) from 689-802aa and those of the Spirochaetes (TC# 3.A.3.27.2) from 814-821aa.

It should be noted that FUPA 27s & 29s have been posited to share function based on their shared phylogeny and topology (Chan, et al. 2010). SEED analysis supports this, as FUPA29 is only found in δ-P and flavobacteria, it is suggested here that 27 and 29 share a common origin in the Proteobacteria common ancestor, a specialized copper transporter. Perhaps after divergence of δ-P, FUPA29 may have been horizontally gene tranferred to Flavobacteria. This model is complicated by the β-Ps Ralstonia solanacearum and Aromatoleum aromaticum, as will be discussed below. All of the operons but the β-Ps' and the ε-Ps' are highly conserved and as such will not be described entirely separately. After the variations found with SEED are noted for each organism in a class, that class's RegPredict analysis will follow.

Two different operon configurations for very similar clusters are seen here, one in which the FUPA27 is typically the 5th-7th gene transcribed in their cco/fix operons, such as the FUPA27-encoding genomes of the α-Ps, γ-Ps and Spirochaetes described in this section, and one in which it is typically the first gene transcribed, such as the β-Ps *A.aromaticum* and *R.solanacearum*, and each of the FUPA29s. All FUPA29 encoding operons start with the FUPA gene and end in a gene encoding a cycZ-like protein, as does that encoding the FUPA27 Rso4. This also makes Rso4 the only βFUPA27 encoded with the cycZ-like protein, which is otherwise restricted to only all γ-Ps except Tcr6. For these notable exceptions, the β-Ps as well as those found in the two additional ε-Ps are discussed and described after the general FUPA27 operons.

The FUPA27 encoding operon typically begins and proceeds as follows (*note: CcoN and CcoO are encoded as separate proteins for the FUPA27 encoding operons, even in the case of Rso4, whereas they are encoded as a single fusion protein in the FUPA29 encoding operons):

1) Cytochrome c oxidase subunit CcoN (EC 1.9.3.1)
2) Cytochrome c oxidase subunit CcoO (EC 1.9.3.1)
3) Cytochrome c oxidase subunit CcoQ (EC 1.9.3.1)
4) Cytochrome c oxidase subunit CcoP (EC 1.9.3.1)
5) Type cbb3 cytochrome oxidase biogenesis protein CcoG, involved in Cu oxidation
6) Type cbb3 cytochrome oxidase biogenesis protein CcoH
7) FUPA27

8) Type cbb3 cytochrome oxidase biogenesis protein CcoS, involved in heme b insertion

*For the (EC 1.9.3.1), $2Cu^{2+}$, which are defined as visible, CuA, and invisible, CuB (characterized by their electron spin properties) are located in subunit II serving in binding of O2 and reduction, and in subunit I serving in electron flow from ferrocytochrome c to the binuclear center, respectively (Ludwig B, 1980; Azzi A, et al., 1990; Chan SI, et al., 1990].

This configuration may be referred to as the CcoNOQPGH-27-S or "knocked-up [en]'gaged 27s".

The α-Proteobacteria:

Atu3 (*Agrobacterium tumefaciens* str. C58 763aa, gi:15888852) has 3 genes transcribed divergently from *ccoG*, interrupting the usual *ccoNOQPGH-27-S* operon. They encode: hypothetical protein (no significant homology), Glyoxalase Family Protein, and Glyoxalase/bleomycin resistance protein/dioxygenase. Divergently transcribed from *ccoN*, 256bp away, a 3-gene operon encoding sulfate permease family protein, a 206aa hypothetical protein homologous to putative ParB-like nuclease and K^+ efflux system KefA protein/Small-conductance mechanosensitive channel.

Also, *ccoS* is followed codirectionally by a gene encoding 6-phosphogluconate dehydrogenase, decarboxylating (EC 1.1.1.44) (may have a Mg^{2+} requirement [Leyva LA, et al., 2008]). Sixty-seven base pairs downstream of this is encoded a 109aa, 2TMS

hypothetical protein with no significant homology which, according to SEED, is not found frequently adjacent to the cco operon and thus is not predicted to be expressed with it.

Bja4 (*Bradyrhizobium japonicum* USDA 110 730aa, gi:27377880) is encoded in the standard *ccoNOQPGH-27-S* operon, with a gene encoding a CBS domain containing membrane protein 210bp and one encoding a UspA domain containing protein 39bp upstream of that. Divergently transcribed from that, 298bp away, a 4 gene-operon encoding two-component oxygen-sensor histidine kinase FixL, two-component nitrogen fixation transcriptional regulator FixJ, transcriptional regulatory protein FixJ (second component) and a transcriptional regulator in the Cpr/Fnr family. At the other end of the operon, the next gene, encoding a 400 residue, 12 TMS protein of the MFS superfamily, is transcribed convergently to *ccoS* and according to SEED, is not found frequently adjacent to the operon; thus it is not predicted to be expressed with it.

Bme4 (*Brucella melitensis* 16M 752aa, gi:17987852) is encoded within an operon which follows the *ccoNOQPGH-27*-S standard pattern. Upstream of *ccoN* by 218bp and codirectionally transcribed an iron response/ferric uptake regulator, of the IRR/FUR family is encoded, however it occurs adjacent to the *cco* operon at very low frequency. A gene transcribed divergently from this, overlapping by 1bp, encodes a 107aa hypothetical protein with no significant homology and 133bp later a gene encoding DsbA family, Com-1 subfamily protein. The next gene, encoding

Salicylaldehyde dehydrogenase, DoxF-like protein, is transcribed convergently and thus is not assumed to be co-expressed. Downstream of ccoS, the next gene is transcribed convergently as well, with a 19bp overlap. It encodes D-glycerate 2-kinase (EC 2.7.1.-) and is not found to occur frequently with the cco operon.

Ccr1 (*Caulobacter crescentus* CB15 724aa, gi:16125656) is encoded within a standard *ccoNOQPGH-27-S* operon which has one gene transcribed divergently upstream of it and 3 genes transcribed codirectionally downstream of it. The upstream gene overlaps *ccoN* by 15bp and encodes a 134aa, 2 TMS hypothetical protein with no significant homology. The downstream genes start 131bp after ccoS with one encoding Outer membrane protein W (ompW) precursor followed 182bp later by one encoding transcriptional activator FtrB of the Crp/Fnr family and finally, 45bp after that is encoded Coproporphyrinogen III oxidase, oxygen-independent (EC 1.3.99.22) (requires Fe^{2+} for a (4Fe-4S) cluster contained within its HemN domain [Layer G, et al., 2005]).

Mlo3 and **Mlo4** both are encoded in complete, conserved *ccoNOQPGH-27-S* operons. The duplicate gene products are designated Cco/Fix(NOQPGHIS)1 and Cco/Fix(NOQPGHIS)2, where “I” refers to the FUPA27, respectively.

The *ccoN* beginning the operon encoding **Mlo3** (*M.loti* MAFF303099 762aa, gi:13475370, fig|266835.1.peg.5035) has several familiar proteins encoded upstream of it, starting with a putative hemerythrin HHE cation binding domain-containing protein encoded divergently, 130bp away. The next genes moving upstream from *ccoN* is worth

mentioning as they have appeared several times before and are shown by SEED to occur frequently near the FUPA27-encoding gene. They encode an Fnr-type transcriptional regulator of the Crp/Fnr family transcribed convergently to the hemerythrin domain encoding gene, followed again divergently, 158bp later, by a gene encoding Coproporphyrinogen III oxidase. The divergent-convergent-divergent nature of the gene cluster makes co-expression unlikely, but recurrence of the same genes as previous mentioned makes them worth noting. At the other end of the operon, downstream of the ccoS gene, 412bp away, a gene encoding a symbiosis island integrase is convergently encoded. Eighty-three base pairs upstream of this gene is one codirectionally transcribed encoding a putative Fic family protein. Due to the distance from the *ccoNOQPGH-27-S* operon and the very low frequency of occurrence near the operon according to SEED however, these gene products are not considered to relate to that of the operon.

Mlo4 (*M.loti* MAFF303099 762aa, gi:13475529, fig|266835.1.peg.5194) is encoded in an operon some 183kb away from that encoding Mlo3. The upstream genetic neighborhood is exactly the same for the operon encoding Mlo4 as that encoding Mlo3, and thus will not be discussed. The downstream neighborhood is quite different, although still seemingly unrelated. The next gene transcribed downstream of the the ccoS for this operon is found 410bp downstream, transcribed convergently. It encodes 232aa, 5 TMS, nodulin 21-related protein. The distance and orientation of the gene, as well as SEED analysis indicating its low occurrence near the operon, imply that the genes are not co-expressed.

Rpa4 (*Rhodopseudomonas palustris* CGA009 732aa, gi:39933093) is encoded in an operon that follows the standard structure mentioned above. *ccoS* is followed by two genes encoding hypothetical proteins of lengths 83 and 70aa co-directionally transcribed 2bp downstream of ccoS and the second 139bp after that. Another 140bp after that is encoded a transcriptional regulator of the AsnC family followed 394bp later by genes encoding Circadian clock protein KaiC and Circadian oscillation regulator KaiB, whose genes overlap by 10bp, then 3bp later a gene encoding Chemotaxis protein methyltransferase CheR (EC 2.1.1.80) and 46bp later this is met convergently by another gene encoding an AsnC family transcriptional regulator. While it is most likely that the large space preceding *kaiC* precludes the possibility of it and later genes being coexpressed with the *ccoNOQPGH-27-S* operon, it should not be completely ruled out. Upstream of *ccoN*, three genes are transcribed co-directionally: 2-dehydropantoate 2-reductase (EC 1.1.1.169) (Mg^{2+} or other divalent cations not required [Matak-Vinkovic D, et al., 2001]) is encoded 311bp away, 73bp upstream of that short-chain dehydrogenase, associated with 2-hydroxychromene-2-carboxylate isomerase family protein is encoded and 121bp upstream of that is encoded the glutathione-dependent 2-hydroxychromene-2-carboxylate isomerase family protein. One gene is transcribed divergently from that, 123bp away, which encodes a Glutathione S-transferase (EC 2.5.1.18) (no known cation requirements in bacteria). The next gene after that is transcribed convergently and thus is assumed not to be coexpressed with the *ccoNOQPGH-27-S* operon. It encodes Short-chain dehydrogenase/reductase (SDR) which may still work in conjunction with the other glutathione genes described above.

Sme5 is encoded in an operon which is missing the *ccoGH* genes and has a less frequently occurring ccoI gene in their place. **Sme6** is encoded in a standard *ccoNOQPGH-27-S* operon. Additionally, this genome encodes a separate CcoNOQP complex found alone. The multiple gene products are designated by MicrobesOnline, RegPredict and others as Cco/Fix(NOQP)3 plus Cco/Fix(IS)2, Cco/Fix(NOQPIS)1 with the only CcoGH and Cco/Fix(NOQP)2, respectively.

Sme5 (*S.meliloti* 1021 755aa, gi:16262778, fig|266834.1.peg.325) is encoded in an operon encoding a slightly different from normal cytochrome c oxidase battery, instead encoding *ccoNOQPI-27-S*. Following *ccoS*, 432bp later, a 54aa, single-TMS hypothetical protein is encoded. The next protein encoded downstream of this, whose gene is co-directionally transcribed, is one of 420aa containing an OmpA domain and a LolA domain, but without any further characterization and encoded 523bp downstream of the hypothetical protein gene, it is most likely not co-expressed with the operon of interest. Upstream of the operon, 4 or 5 genes are found transcribed co-directionally, depending on whether or not two are actually both coding fragments of a single larger gene that is more commonly found or a sequencing error occurred and it really is just one gene. In order of transcription they encode:

1) another protein similar to that mentioned above found downstream but containing only the LolA domain
2) -3bp later (overlap) a 111aa hypothetical protein with no significant homology

3) 53bp later, a 517aa, 1 TMS putative exported protein with no significant homology containing DUF1254 and DUF1214 superfamily domains
4) 58bp later, the first of two genes that may actually be a missequenced single gene encodes a 192aa protein containing a DUF1254 domain
5) 6bp later, the second gene encodes a 287aa protein containing a DUF1214 domain.

The *ccoN* gene is found 305bp downstream of this. Thus, while coexpression is not likely, it may be possible. The next gene upstream of that encoding (1) LolA domain protein however is 550bp away and does not occur with any frequency near the cco operon, so it is not considered likely to be coexpressed.

Sme6 (*S.meliloti* 1021 757aa, gi:16263112, fig|266834.1.peg.659) is encoded within a regular *cco/fixNOQPGH-27-S* operon, and more extensive coexpression with operons is suggested by their arrangement. Five genes are found upstream of the *cco/fixNOQPGH-27-S* operon whose expression is at the least related if not co-regulated with that of the *cco/fixNOQPGH-27-S* operon. In order of transcription they encode:

1) Two-component oxygen-sensor histidine kinase FixL
2) 7bp later, Two-component nitrogen fixation transcriptional regulator FixJ
3) 99bp later, FixT2 transcription regulator (5 FixJ/K promoters occur between this and the next gene)
4) 80bp later, transcriptional regulator, Crp/Fnr family family protein
5) 81bp later, Pyridoxamine 5'-phosphate oxidase-related, FMN-binding protein

*239bp after this a putative FixK site (FNR box) at *fixN* gene (15bp long) occurs, followed 61bp later by the *cco/fixN* gene.

Transcribed divergently from the *fixL* gene is a single gene encoding an universal stress protein An (UspA) gene 234bp away which might be co-expressed. Downstream of this gene is an operon encoding NapEFDC and other proteins involved in NO3- and NO2- ammonification, but since they cannot share a promoter, they are not assumed to be co-expressed.

Downstream of the *cco/fixNOQPGH-27-S* operon, 8 genes are predicted to exist, starting 53bp after *cco/fixS*. In order of transcription, they encode:

1) transcriptional regulator Crp/Fnr protein (see #4 above)
2) 237bp later, a 126aa hypothetical protein with 1 TMS and no significant homology
3) 255bp later, nuclear export factor GLE1
4) 65bp later, Copper resistance protein D
5) 54bp later, a 120aa hypothetical protein with no significant homology
6) 59bp later, Nitric oxide-dependent regulator DnrN or NorA
7) 16bp later, NnrS protein involved in response to NO
8) 169bp later, a 57aa hypothetical protein with no significant homology

Transcribed convergently to (8), 79bp away, is a gene encoding Flavohemoprotein (Hemoglobin-like protein) (Flavohemoglobin) (Nitric oxide dioxygenase) (EC 1.14.12.17) (no known metal requirements), and upstream of that the NosRDFYLX, involved in denitrification, is encoded.

Spo3 (*Silicibacter pomeroyi* DSS-3 725aa, gi:56698342) is encoded in a normal *cco/fixNOQPGH-27-S* operon. Divergently transcribed from the *ccoN* gene is a single gene encoding a universal stress protein UspA. After this gene the next gene is transcribed convergently to it, encoding a hypothetical membrane protein, is presumed not to be co-expressed. Downstream of the operon are small genes encoding (in order of transcription):

1) 48bp afer *ccoS*, a 141aa Transcriptional regulator in the Cupin 2 superfamily
2) 23bp later, Preprotein translocase subunit SecE (TC 3.A.5.1.1)
3) 203bp later, Transcription antitermination protein NusG

Convergently transcribed to this is a gene encoding iron-regulated protein frpC, which is also presumed not to be coexpressed.

RegPredict Analysis: co-regulation of *cco/fixNOQPGH-27-S* with non-local genes in α-Proteobacteria shows a pattern of further involvement of the operon, and this FUPA27, also sometimes known as CcoI, with cytochrome c and its affiliated proteins. One likely regulatory sequence motif preceding the operon in α-Proteobacteria (CGCnTTGATnNNnATCAAnGCG) is found preceding nearly every *ccoN*, including 2 copies each in *S.meliloti* and *M.loti* and in 6 cases, also preceding *ccoG*. The specific sequences are as follows; listed by the gene they precede:

Agrobacterium tumefaciens *ccoN* tttCTTGATCt|gGATCAAGgtg, score=5.62

	ccoG	tcaCTTGAtgT|AagTCAAGgat, score=5.21
Bradyrhizobium japonicum	*ccoN*	tatcTTGATTt|cAATCAAttcc, score=5.19
	ccoG	cCgtTTGAgCt|gGaTCAAcgGa, score=4.40
Brucella melitensis	*ccoN*	tTgtTTGATtT|AgATCAAtgAc, score=5.48
Caulobacter crescentus	*ccoN*	GGCTTTGAtgC|GtgTCAAAGCC, score=5.43
Mesorhizobium loti	*ccoN*1	gcaCTTGATCt|gGATCAAGgtg, score=5.23
	*ccoG*1	gcgaTTGacct|gcagCAAggcg, score=4.20
	*ccoN*2	gGCtTTGAcCt|gGaTCAAgGCg, score=6.06
	*ccoG*2	gGgaTTGAcct|gcaTCAAcgCg, score=4.52
R.palustris	*ccoN*	catTTTGATtt|cgATCAAAcgt, score=5.03
Sinorhizobium meliloti	*ccoN*1	gcaCTTGATCt|gGATCAAGgtg, score=5.66
	ccoG	agaCTTGAcgc|agaTCAAGgtg, score=5.02
	*ccoN*2	tgtCTTGATct|gaATCAAGgtg, score=5.67
	*ccoN*3	(none found above threshold)
Silicibacter pomeroyi	*ccoN*	CGtcTTGATgc|agATCAAagCG, score=5.81

ccoG CttTTTGAccc|ataTCAAAgtG, score=4.70

Related sequences found in these genomes precede and potentially regulate several gene clusters mostly involved in cytochrome c activity. One such cluster precedes the *cycHJKL* operon in *R.palustris*, *A.tumefaciens* and *M.loti*. This operon encodes the proteins:

cytochrome c heme lyase subunit, CcmH

cytochrome c-type biogenesis protein/heme chaperone, CcmE

cytochrome c heme lyase subunit, CcmF

cytochrome c heme lyase subunit, CcmL

In the case of the latter two, the gene dop follows these and encodes a serine DO-like protease. The specific sequences are as follows, all preceding *cycH*:

R.palustris	ccctTTGATTc|aAATCAAgttt, score=4.81
A.tumefaciens	cAgtTTGAcgT|AaaTCAAggTc, score=4.96
M.loti	ccggTTGATgt|cgATCAAggac, score=4.95

One other genome, *S.pomeroyi*, contains a related potential regulatory sequence (CcgCcTGACtt|cgGTCAaGgcG, score=4.88) in this set, however the 3 genes it is predicted to regulate are not as well defined. The second gene still encodes CcmE here, but the one preceding it, *argC*-2, encodes N-acetyl-gamma-glutamyl-phosphate

reductase (EC 1.2.1.38) (no known metal ion requirements, though at least partial divalent cation inhibition is found [Vogel HJ; 1970]) and the one following it encodes a secretion activator protein containing glycosyl hydrolase 108 superfamily and peptidoglycan binding 3 superfamily domains.

The next set in this regulatory motif is predicted to regulate expression of genes less closely related to the *cco/fix* operon. The first (CgtTTTGcttt|cggtCAAAgtG, score=5.04), found in *B.melitensis*, is predicted to regulate an operon encoding adenosine (5')-pentaphospho-(5")-adenosine pyrophosphohydrolase (EC 3.6.1.-), then a 3 TMS transglycosylase associated protein and then an 8 TMS, 215aa putative membrane protein. In *S.meliloti*, a similarly related sequence (ggCcTTGtCtt|cgGgCAAaGag, score=4.61) precedes a much larger operon that encodes the following 14 proteins in order of transcription (gene names included where available):

1) *rplU* encodes LSU ribosomal protein L21p
2) *rpmA* encodes LSU ribosomal protein L27p
3) 50S ribosomal protein acetyltransferase
4) 50S ribosomal protein acetyltransferase
5) *obgE* encodes GTPase ObgE
6) *proB1* encodes glutamate 5-kinase (EC 2.7.2.11) (forms dimers requiring two Mg^{2+} each [Marco-Marin C, et al., 2007])
7) *proA* encodes gamma-glutamyl phosphate reductase (EC 1.2.1.41) (strongly inhibited by divalent cations including Cu^{2+} [Hayzer DJ, et al., 1982])
8) *nadD* encodes nicotinic acid mononucleotide adenylyltransferase

9) Iojap protein
10) rRNA large subunit methyltransferase
11) Peptidase M23B precursor
12) *ctpA* encodes carboxyl-terminal protease (EC 3.4.21.102) (Ca^{2+}, Co^{2+} and Mn^{2+} cause activation in *E.coli* [Silber KR, et al.,1992])
13) possible divergent polysaccharide deacetylase
14) i*alA* encodes adenosine (5')-pentaphospho-(5")-adenosine pyrophosphohydrolase

Three of these genes appear as a smaller operon regulated by a related sequence (cGCCTTGgTtT|AgAgCAAGGCc, score=5.07) in *M.loti*. These genes are:

1) *obgE* which encodes a GTP-binding/GTPase protein ObgE
2) *proB1* which encodes glutamate 5-kinase (EC 2.7.2.11)
3) *proA* which encodes gamma-glutamyl phosphate reductase (EC 1.2.1.41)

The next set in this motif precedes operons largely relating to cytochrome c oxidase genes. The first two are both found in *S.pomeroyi*. One (tAtcTTGATCc|aGATCAAacTc, score=5.01) precedes a five-gene operon encoding the following proteins in order of transcription:

1) *ctaD* encodes cytochrome c oxidase polypeptide I (EC 1.9.3.1)
2) 308aa hypothetical protein with a glycosyltransferase family A/2 domain
3) 158aa, 2TMS hypothetical protein in the DUF2244 superfamily, unknown function

4) 165aa, 1TMS hypothetical protein identified by similarity to the *M.loti* protein with accession #: GB:BAB50627.1. Neither of these protein have significant homology to any known protein
5) prephenate and/or arogenate dehydrogenase (unknown specificity) (EC 1.3.1.12) (EC 1.3.1.43), NAD-specific (no unique metal ion requirements are found for either EC #).

The second member of this set (tTgTTTGATCc|aGATCAAAaAt, score=4.60) has more in common with the rest of the set in terms of which genes are predicted to be regulated. It precedes a 10-gene operon that encodes the following in order of transcription:

1) *ctaC* encodes cytochrome c oxidase polypeptide II (EC 1.9.3.1)
2) *cyoE/ctaB* encodes heme O synthase, protoheme IX farnesyltransferase (EC 2.5.1.-) COX10-CtaB
3) *ctaG* encodes cytochrome oxidase biogenesis protein Cox11-CtaG, copper delivery to Cox1
4) *ctaE* encodes cytochrome c oxidase polypeptide III (EC 1.9.3.1)
5) cytochrome oxidase biogenesis protein Surf1, facilitates heme A insertion
6) *thrC* encodes threonine synthase (EC 4.2.3.1) (no known metal requirements)
7) *mpp* encodes mitochondrial processing peptidase-like protein, M16 family (EC 3.4.24.64) (no metal requirement data available for prokaryotes)
8) *rimJ* encodes ribosomal-protein-S5p-alanine acetyltransferase

9) a 282aa hypothetical protein with 56% identity to beta-lactamase fold-like Zn-dependent hydrolase
10) D-2-hydroxyglutarate dehydrogenase

The next member of this set (gGccTTGAtCg|gGcTCAAatCg, score=4.63), found in *C.crescentus*, precedes a 12-gene operon encoding, in order of transciption:

1) *ctaC* encodes cytochrome c oxidase polypeptide II (EC 1.9.3.1)
2) *ctaD* encodes cytochrome c oxidase polypeptide I (EC 1.9.3.1)
3) *cyoE/ctaB* encodes heme O synthase, protoheme IX farnesyltransferase (EC 2.5.1.-) COX10-CtaB
4) 53aa, 1 TMS hypothetical protein 44% identical to cytochrome C oxidase assembly protein CoxF *in Oceanicaulis alexandrii*
5) *ctaG* encodes cytochrome oxidase biogenesis protein Cox11-CtaG, copper delivery to Cox1
6) *ctaE* encodes cytochrome c oxidase polypeptide III (EC 1.9.3.1)
7) 120aa, 2TMS conserved hypothetical protein in cytochrome c oxidase gene clusters with homology to cytochrome c oxidase subunit III zinc finger motifs
8) cytochrome oxidase biogenesis protein Surf1, facilitates heme A insertion

9.1) *thrC* encodes threonine synthase (EC 4.2.3.1)

10.1) *mpp* encodes mitochondrial processing peptidase-like protein (EC 3.4.24.64), M16 family

11.1) *rimJ* encodes ribosomal-protein-S5p-alanine acetyltransferase

12.1) a diguanylate cyclase (GGDEF) family protein

The last member of this set (GgGtTTGATCc|tGATCAAgCgC, score=4.64), found in *S.meliloti*, precedes another similar 11-gene operon identical to that of *C.crescentus* for the first 8 genes, but then followed instead with:

9.2) *lytB* encodes 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (EC 1.17.1.2) (Co^{2+} and Mn^{2+} act as activators at 0.5 mM and enzyme possesses a dioxygen-sensitive [4Fe-4S] cluster, but these cations, as well as Ca^{2+}, Mg^{2+}, Ni^{2+} and Zn^{2+} act as inhibitors at higher concentrations [Wolff M, et al., 2003; Graewert T, et al., 2004])

10.2) *thrB* encodes homoserine kinase (EC 2.7.1.39) (Mg^{2+} is absolutely required, optimal activity between 10 mM and 20 mM in *E.coli* [Théze J, et al., 1974])

11.2) *rnhA1* encodes ribonuclease H I (EC 3.1.26.4) (requires Mg^{2+} or Mn^{2+} to function, and in presence of 10 mM of Ba^{2+}, Ca^{2+}, Co^{2+}, Zn^{2+}, Cu^{2+}, Fe^{2+}, or Sr^{2+} [Ohtani N, et al., 2000]).

The next member of this set is found in 3 species, all regulating the gene for outer membrane W precursor in all 3 cases. The 3 species are *B.japonicum*, *R.palustris* and *B.melitensis* and their sequences (GGATTTGAtcg|gcgTCAAATCC, score=4.23; tcATTTGATCc|aGATCAAATct, score=5.14; CtCtTTGATtt|ggATCAAtGgG, score-5.55, respectively) either precede this single gene alone or, in the case of *R.palustris*,

precede the outer membrane protein W precursor gene and a second gene encoding alkylphosphonate utilization operon protein PhnA as well.

The next members of this set are found in *S.meliloti* and *M.loti* and may represent two different ways of achieving enhanced expression. They both precede genes for universal stress protein UspA, but interestingly, whereas *S.meliloti* encodes 2 copies of the gene preceded by a single regulatory sequence (TttcTTGAcCg|gGaTCAAtgtA, score=4.54), *M.loti* encodes only 1 copy preceded by two copies of the predicted regulatory sequence (tgcgTTGATgC|GgATCAAatgc, score=4.88; gTgtTTGattc|agcgCAAtgAa, score=4.15).

The last members of this set are found in *B.japonicum* and *A.tumefaciens*. In *A.tumefaciens*, the sequence (cggtgTGATCa|aGATCAaggaa, score=4.51) precedes a 4-gene operon encoding:

1) *napD* encodes periplasmic nitrate reductase (PNR), NapD protein
2) *napA* encodes PNR large subunit (EC 1.7.99.4)
3) *napB* encodes PNR cytochrome c550-type small subunit
4) *napC* encodes PNR, cytochrome c-type protein

The *B.japonicum* sequence (CGgaTTGATCc|aGATCAAcgCG, score=4.88) precedes a 6-gene operon identical to this except that it has an additional protein encoded at each end of the operon. At the beginning of the operon napE encodes periplasmic nitrate reductase component NapE and at the end of the operon *edrN* encodes exodeoxyribosylnuclease/recD-like DNA helicase.

The γ-Proteobacteria:

*All of the γ-P genomes encode cycZ after CcoS, except that of *Thiomicrospira crunogena*, and although the order is maintained, a "putative CcoH" sequentially distinct enough not to identified as homolgous by SEED's PSIBLAST function is encoded in the place of the CcoH discussed thus far. This analog is recognized as having the essential CcoH/FixH domain necessary to be considered such according to CDD.

Cps2 (*Colwellia psychrerythraea* 34H 820aa, gi:71279318): is encoded in an operon also encoding CcoNOQPH-FUPA27-S-cyc-Z in the usual configuration. However, the operon does not contain *ccoG*, although two *ccoG* genes are found separately, neither of them near the operon. Following the gene encoding the cycZ-like protein are two other frequently occurring genes transcribed co-directionally; the first is 54bp downstream, the other 146bp after that. They encode a fumarate and nitrate reduction regulatory protein (aka, another Crp/Fnr family transcriptional regulator) and a universal stress protein E (UspE) respectively. Downstream of these, no pattern is conserved and genes are transcribed convergently. Upstream of *ccoN*, albeit 432bp away, another Crp/Fnr family transcriptional regulator is encoded. One gene is transcribed divergently, 198bp away and occurring at only a low frequency adjacent to the FUPA27-encoding operon, but its orientation does suggest coexpression. It encodes a 157aa hypothetical protein referred to as a lipoprotein despite no apparent TMSs and no significant homology.

Similary to Cps2, **Ilo3** (*Idiomarina loihiensis* L2TR 792aa, gi:56460410) is encoded in an otherwise standard *ccoNOQPGH-27-S-cycZ* operon except that *ccoG* is absent from the operon and instead one *ccoG* is found separately, no where near the operon. Sixty-nine bp downstream of *cycZ*, a gene encoding universal stress protein E is found, followed by one encoding tRNA (cytosine32)-2-thiocytidine synthetase 67bp after that. The next gene is transcribed convergently, ending 70bp from that encoding the tRNA synthetase, and encodes a BAX protein. Upstream of *ccoN*, 3 genes are transcribed co-directionally, encoding, in order of transcription:

1) probable RND efflux membrane fusion protein
2) -7bp later (overlap), acriflavin resistance protein
3) 100bp later, methyl-accepting chemotaxis sensory transducer- ccoN is encoded 143bp later.

One gene is encoded divergently from that encoding probable RND efflux membrane fusion protein, 103bp away, which encodes a 63aa zinc-ribbon protein. Convergently transcribed to this is a gene encoding a FMN-dependent NADH-azoreductase, which cannot share a regulatory region and is found near the operon only at low frequency.

Three of the four Pseudomonads described in Chan, et al. (2010) have very similar variations on the general structure of the *ccoNOQPGH-27-S-cycZ* operon. The structure of the operon of *P.syringae* more closely resembles the usual configuration.

The variation in the configuration is a duplication of *ccoNOQP* directly upstream of the regular operon (albeit 365-586bp away), making the arrangement into *ccoNOQPNOQPGH-27-S-cycZ*. In all cases of duplication, the set of duplicates transcribed first are denoted as *cco(NOQP)*1 and the second set as *cco(NOQP)*2. Aside from the partial operon duplication, all four Pseudomonas operons are very similar, even including many of the same potentially coexpressed flanking genes. Divergently transcribed from the most upstream of the *ccoN* in all cases is a single gene encoding an esterase/lipase/thioesterase family protein (-7 to 188bp away). Downstream of *cycZ*, transcribed co-directionally 94-149bp away, coproporphyrinogen III oxidase, oxygen-independent (EC 1.3.99.22) is encoded, followed 4-49bp by a gene transcribed convergently which encodes FIG00953282: hypothetical protein, a 156-185aa long protein that has no significant homology in 3 Pseudomonas species but followed instead 187bp later by a co-directional gene encoding fumarate and nitrate reduction regulatory protein (another Crp/FNR family transcriptional regulator) in *P.putida*. The other three Pseudomonas species also encode this Crp/FNR family transcriptional regulator in the same direction as the coproporphyrinogen III oxidase gene 79-246bp from the start of the hypothetical protein. Downstream of that, all four species have the same protein encoded as well; 56-86bp further downstream, co-directionally adenine phosphoribosyltransferase (EC 2.4.2.7) (enzyme requires Mn^{2+} or Mg^{2+} [Hochstadt J, 1978]). The frequency of the co-directionally transcribed genes adjacent to the cco operon is higher than that of the hypothetical protein and thus they are implicated in co-

expression despite interruption. Below, the notable, relevant differences between the operons' neighborhoods and homologues are briefly described.

Pae4 (*P.aeruginosa* PAO1 811aa, gi:15596746) is encoded within an operon unique to the Pseudomonas species as described above. The only local variations remaining to speak of in this region are which convergently transcribed genes probably signify the end of the transcription unit. In this species, upstream of the esterase/lipase/thioesterase family protein gene and 89bp away, a nucleoside-diphosphate-sugar epimerase is encoded codirectionally but occurs near the operon at low frequency in general. This is followed 38bp later convergently by a gene encoding an aerotaxis sensor receptor protein Aer. Exactly 115bp away, a putative cytoplasmic protein is encoded, and beyond it are ever more infrequently co-localized genes.

Downstream of the extended possible operon, after the gene encoding adenine phosphoribosyltransferase, 3 more proteins are encoded co-directionally. In order of transcription, they are:

1) 240bp downstream, a 2TMS metal-dependent hydrolase-like protein
2) 409bp later, spermidine export protein MdtJ
3) -6bp (overlap) later, spermidine export protein MdtJ (duplicate)

The next gene is convergently transcribed, overlapping by 146bp and encodes a transcriptional regulator of the TetR family. Both the intergenic distance and the low frequency of these genes suggest that they are not expressed with the *cco* operon.

Other related characteristics unique to this species include two more additional distant copies of *ccoN*: one which is found with an additional *ccoG* and one which is which is found without any other *cco* genes.

Pfl4 (*P.fluorescens* Pf-5 778aa, gi:70729297) is also encoded in an operon with only local variations remaining to speak of in this region which are convergently transcribed genes probably signifying the end of the transcription unit. In this species, upstream of the esterase/lipase/thioesterase family protein gene and 168bp away, the aerotaxis sensor receptor protein Aer is also encoded convergenlty and beyond it are ever more infrequently co-localized genes. The next gene downstream of the adenine phosphoribosyltransferase gene is 8bp away and encodes a Butyryl-CoA dehydrogenase (EC 1.3.99.2) (no known metal requirements) convergently. A lone third copy of *ccoN* also found.

Ppu3 (*P.putida* KT2440 882aa, gi:26990952) is the third and final FUPA27 homologue encoded in an abnormal Pseudomonas *cco* operon. In this species, upstream of the esterase/lipase/thioesterase family protein gene and 115bp away, a putative cytoplasmic protein is encoded convergently and beyond it are ever more infrequently co-localized genes. A lone second copy of *ccoG* is also found.

Psy2 (*P.syringae pv. tomato str*. DC3000 823aa, gi:28869200) is encoded in an operon which, in addition to being the only *cco* operon in Pseudomonas without a duplication of the *ccoNOQP* portion of the operon, has two addition genes found between *ccoP* and *ccoG*. They are codirectional to the normal operon and encode a toxin/antitoxin system consisting of VapBC (VapC is the toxin) and noted as separate

from the RelBE and MazEF systems. Upstream of the esterase/lipase/thioesterase family protein gene and 2bp away, a 158aa hypothetical protein is encoded convergently. The next gene downstream of the adenine phosphoribosyltransferase gene is 131bp away and encodes a 160aa hypothetical protein convergently which is 69% identical to a methyl-accepting chemotaxis protein (gi:379063492). There are no cco genes found outside of the operon in this genome.

Thus concludes the FUPA27 colocalization analysis for the Pseudomonads.

Pha2 (*Pseudoalteromonas haloplanktis* TAC125 791aa, gi:77360787) is encoded within a fairly normal *cco/fix* operon but for the notable exception that the operon is lacking *ccoG*. Two copies of the gene are found elsewhere in the genome however, without any other members of the operon. In regards to co-localized genes which may be coexpressed with the operon, there are 3 divergently transcribed from *ccoN* and 7 downstream and co-directional from *cycZ*. In the order of transcription, the genes transcribed divergently from ccoN begin 344bp away with one encoding a 140aa, 1 TMS hypothetical protein with no significant homology, followed 823bp later by a gene encoding TonB-dependent receptor and 105bp after that by a gene encoding a 455aa, 2 TMS hypothetical protein with no significant homology. The next gene after this is transcribed convergently 51bp away and encodes a putative lipoprotein. Both the intergenic distance and the low frequency of colocalization of these genes make their co-expression with the cco operon unlikely, however. Downstream of *cycZ*, the first 3 of the

7 genes codirectionally transcribed are both close and frequently co-localized. The proteins they encode are listed below in order of transcription:

1) 42bp after *cycZ* comes fumarate and nitrate reduction regulatory protein (Crp/FNR family)
2) 153bp later, universal stress protein E, UspE
3) 88bp later, tRNA(Cytosine32)-2-thiocytidine synthetase
4) 206 putative orphan protein (42aa, 0 TMS, may not be expressed at all)
5) 71bp later, FIG002283: Isochorismatase (cysteine hydrolase) family protein
6) 52bp later, Cytochrome c553
7) 21bp later, Exoenzymes regulatory protein AepA precursor

The next gene after that is transcribed convergently 137bp away and encodes RecD, Exodeoxyribonuclease V alpha chain (EC 3.1.11.5), the last part of the RecBCD DNA repair pathway. It is not likely that 4-7 serve to function with the cco operon given their low frequency of colocalization.

Ppr2 (*Photobacterium profundum* SS9 769aa, gi:54309022) is encoded in a *cco* operon in which *ccoG* is absent from operon but is instead found alone elsewhere in the genome. The operon is otherwise in the standard configuration. Upstream of *ccoN* and divergently transcribed 272bp away is encoded a 147aa, 1 TMS hypothetical protein with no significant homology, followed 151bp later by a convergently transcribed gene encoding oxidoreductase, short chain dehydrogenase/reductase family. While the hypothetical protein is of an intergenic distance from ccoN that it may be coexpressed,

it's frequency near the cco operon is so low that co-functionality seems unlikely. Downstream of *cycZ*, two familiar genes are codirectionally transcibed: 75bp away is one encoding fumarate and nitrate reduction regulatory protein (Crp/FNR family) and 194bp after that is one encoding UspE. As previously mentioned, these both occur near the cco operon at high frequency. Convergently transcribed from *uspE*, 114bp away is encoded a 117bp hypothetical protein with no significant homology and 411bp divergently transcribed from that (and thus codrectional to the cco operon) is encoded the tRNA(Cytosine32)-2-thiocytidine synthetase mentioned previously, although how it could be transcribed directly with the operon in this case is unclear.

Son2 (*Shewanella oneidensis* MR-1 799aa, gi:24373906) is encoded in a *cco* operon in which *ccoG* is absent from the operon but is instead found twice alone elsewhere in the genome. The operon is otherwise in the standard configuration. Upstream of *ccoN* and divergently transcribed 226bp away is encoded a 161aa, 1 TMS hypothetical protein with no significant homology. Downstream of *cycZ*, three frequently occurring proteins are encoded starting 44bp downstream with fumarate and nitrate reduction regulatory protein (Crp/FNR family) followed 127bp after that is encoded UspE and 114bp after that is encoded tRNA(Cytosine32)-2-thiocytidine synthetase. The next gene is again convergently transcribed and encodes a 207aa hypothetical protein in the DUF2987 superfamily that appears to be restricted to the γ-Proteobacteria but has no significant homology.

Tcr6 (*Thiomicrospira crunogena* XCL-2 828aa, gi:78486299) is encoded in a *cco/fix* operon in which *ccoQ* is absent and not found anywhere else in the genome. The *ccoH* identified in this organism appears to be more closely related to those in the α-Proteobacteria than to the homologues otherwise described for γ-Proteobacteria. Additionally, a separate, single 4TMS protein-encoding *ccoI* is found elsewhere in the genome whose product appears to bear no homology to the FUPA27 which is also referred to as CcoI. As for the gene neighborhood, that downstream of the operon is most likely not coexpressed as the very next gene after *cycZ* is convergently transcribed and encodes the small chain of a glutamate synthase [NADPH] (EC 1.4.1.13) which clearly occurs near the cco operon at sparingly low frequency. The upstream neighborhood however contains 11 co-directional genes starting from one divergently transcribed 395bp from *ccoN*, encoding proteins involved in histidine biosynthesis and the twin arginine translocation system. The proteins encoded are, in order of transcription:

1) imidazoleglycerol-phosphate (IGP) dehydratase (EC 4.2.1.19) (stabilized by Mn^{2+} [Brady DR, et al., 1973])
2) 4bp later, IGP synthase amidotransferase subunit (EC 2.4.2.-)
3) 17bp later, phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase (EC 5.3.1.16) (no known metal requirements)
4) 10bp later, IGP synthase cyclase subunit (EC 4.1.3.-)
5) -3bp later (overlap), Phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) (requires Mg^{2+} [Javid-Majd F, et al., 2008])
6) 21bp later, HIT family hydrolase

7) 129bp later, twin-arginine translocation (Tat) protein TatA
8) 46bp later, TatB
9) -3bp later (overlap), TatC
10) 23bp later, thymidylate kinase (EC 2.7.4.9) ("absolute requirement for divalent cation. When Mg^{2+} is equal to ATP, the rate of dTMP kinase reaction is maximal", additionally Mn^{2+} and Co^{2+} may substitute for Mg^{2+}, with 41% and 18% of the activity acheived with Mg^{2+}, respectively [Nelson DJ, et al., 1969])
11) 4bp later, probable zinc-dependent metallopeptidase

The next gene after this is transcribed convergently 7bp away, encoding competence protein F homologue, phosphoribosyltransferase domain/protein YhgH required for utilization of DNA as sole source of carbon and energy. This gene is clearly not cotranscribed and since it and those genes following it are only found at very low frequency near the cco operon and the other genes just listed; it is treated as the end of the possible extended operon.

The three FUPA27s in Vibrio species described in Chan, et al. (2010) are all encoded in *cco* operons which follow the same exact pattern: *ccoG* is absent from operon but found elsewhere and the configurations are otherwise normal. *V.parahaemolyticus* and *V.vulnificus* also have ccoI genes elsewhere in their genomes, although again, these bear no resemblance to the FUPA27 gene. The *cco* operons of all 3 genomes also have 3 familiar genes immediately downstream of them:

1) 87-90bp after *cycZ* comes fumarate and nitrate reduction regulatory protein (Crp/FNR family)

2) 130-180bp later, universal stress protein E, UspE

3) 106-124bp later, tRNA(Cytosine32)-2-thiocytidine synthetase

The one exception to this is that in *V.parahaemolyticus*, #3 is convergently transcribed. In the other two Vibrio species, it is co-directional and the operon is most likely ended instead after this because in both species the tRNA synthetase encoding gene is convergent with one encoding hypothetical protein, bax gene locus, 70-91bp away.

Upstream, all 3 operons share the same 3-gene divergently transcribed operon; the first two genes encode hypothetical proteins, the third a histidine kinase. The encoded proteins, in order of transcription, are:

1) 221 to 344bp away, 89-128aa, 1 TMS hypothetical protein with no significant homology
2) 69 to 116bp away, a 393-405aa hypothetical protein containing FISTN and FISTC domains (at the N and C termini, respectively) which are suggested to bind small ligands like amino acids
3) -3 to -13bp away (overlap), signal transduction histidine kinase/response regulator.

The only appreciable difference between the Vibrio *cco* operons seems to be in which gene is transcribed convergently from the 3-gene operon divergently transcribed and possibly coexpressed which is described above. For *V.cholerae,* that gene is 61bp

away and encodes a RTX-I toxin secretion ATP-binding protein RtxE, one end of a bidirectional Rtx toxin-handling transcription unit. For *V.parahaemolyticus*, it is a gene 168bp away encoding bacteriophage237 ORF10. Finally, for *V.vulnificus*, the convergent gene is 246bp away and encodes a 406aa hypothetical protein with no significant homology.

For reference, the FUPA27s as they are named and identified in Chan, et al. (2010) are as follows:

Vch2 (*V.cholerae O1 biovar El Tor str*. N16961 790aa, gi:15641448)

Vpa2 (*V.parahaemolyticus* RIMD 2210633 787aa, gi:28898313)

Vvu2 (*V.vulnificus* CMCP6 789aa, gi:27365909)

RegPredict Analysis: Co-regulation of *cco/fixNOQPGH-27-S-cycZ* with non-local genes in γ-Proteobacteria shows a pattern of further involvement of the operon, and this FUPA27, also sometimes known as CcoI, with cytochrome c and its affiliated proteins. The genome of *T.crunogena* is not available in RegPredict and thus is not included in this discussion. One likely regulatory sequence motif preceding the operon in γ-Proteobacteria (nnTTGATnT|AnATCAAnn) is found preceding 6 of the *cco* operons just before ccoN or ccoO and 2 others slightly later in the operon instead, sometimes with multiple sequences. Several other genes are also included in this first set as well which as noted above are consistently found near the cco operon. The specific sequences are as follows; listed by the gene they precede:

C.psychrerythraea	*ccoN*	ttaTTGATcT|AcATCAAatt, score=3.87
		GatGTGAcAT|ATaTCACagC, score=4.47
	ccoO	ctCTTgAagT|AgcTtAAGga, score=3.91
I.loihiensis	*ccoN*	gtaTTGAtTt|gAgTCAActt, score=4.28
	hemN	tTtTTGACtg|ggGTCAAgAg, score=4.00
	ccoS	CaTcTGattT|AgcgCAaAcG, score=3.76
P.haloplanktis	*uspE*	agtTTGaTtT|AgAaCAAtaa, score=4.26
	hemN	tttTTGtTgT|AgAtCAAccc, score=4.51
P.fluorescens	*ccoO*1	gCtTcGAtgt|gggTCaAcGg, score=4.75
	*ccoP*1	GaCccGatgc|cgggCacGgC, score=3.67
	*ccoN*2	GCatTGACCc|aGGTCAtaGC, score=4.24
	hemN	ccCTTGAtac|aagTCAAGat, score=4.18
P.aeruginosa	*usp7*	CgtTTGATat|gcATCAAgaG, score=4.12
	usp5	gGacTGAcCc|aGaTCAagCg, score=3.91
	fnr	TgttTGACgc|aaGTCActgA, score=4.15
	hemN	CgcTTGATac|aaATCAAcaG, score=3.97
P.putida	*ccoO*1	gCcTcGAtgt|gggTCaAcGg, score=4.75

	ccoG	CAAgTGAcCt|gGaTCAaTTG, score=3.98
	usp7	tGCTTGATGT|ACATCAAGCg, score=4.33
	hemN	cccTTGATac|aaATCAAcac, score=4.49
P.syringae	*vapC*	GtgctGcTgt|cgAtCgaggC, score=4.01
	ccoG	GgcCtGAcct|gtaTCgGcgC, score=3.73
	hemN	cttTTGATcc|acATCAAggc, score=5.04
P.profundum	*fnr*	AtaTTGAccT|AcaTCAAatT, score=3.82
	uspE	ctAtTGATgC|GtATCAcTtt, score=3.78
S.oneidensis	*ccoN*	AgaTTGATCT|AGATCAActT, score=4.20
		TttTTGAcat|ggcTCAAgtA, score=4.42
		gcaTTGAcaT|AgcTCAAcaa, score=3.66
	uspE	aaCTTGaTTT|AAAaCAAGca, score=4.50
		actTTgcTaT|AaAtaAActa, score=3.74
V.cholerae	*fnr*	atATTGAcGT|ACaTCAATta, score=4.00
V.parahaemolyticus	*fnr*	ataTTGATgT|AtATCAAata, score=4.48
V.vulnificus	*ccoH*	tCTTTccTTT|AAAttAAAGc, score=3.86

Unnamed*	acttTgATat|ggATgAttac, score=3.82

*genes encoding periplasmic/membrane protein associated with DUF414 followed by *hemN*

Related sequences also preceded the gene encoding outer membrane protein W precursor in all genomes except those of *C.psychrerythraea, P.profundum* and *V.vulnificus*. These are described below:

I.loihiensis	agGTTGATgc|agATCAACtc, score=4.39
P.haloplanktis	aacTTaATcT|AaATcAAtaa, score=3.99
	AcCTTGaTcT|AaAaCAAGtT, score=4.32
P.fluorescens	tCaTTGAgct|gagTCAAgGc, score=4.23
P.aeruginosa	TgaTTGAgct|gtgTCAAggA, score=3.79
P.putida	tcgTTGAgCt|gGgTCAAact, score=3.76
P.syringae	tCaTTGAgct|gagTCAAaGt, score=3.90
S.oneidensis	aTtTTGATtt|ggATCAAtAa, score=5.18
V.cholerae	AAaTTGATtt|ccATCAAgTT, score=4.02
V.parahaemolyticus	AaaTTGATtt|ccATCAAaaT, score=3.72

A third set of related sequences in this set precedes a gene in 5 genomes variably referred to as *ccpA* or *ccpR* and encoding cytochrome c551 peroxidase (EC 1.11.1.5):

V.cholerae	tcTTTGaTTT|AAAgCAAAtc, score=4.80

V.parahaemolyticus	aaaTTGAcTt|tAaTCAAgga, score=4.41
C.psychrerythraea	gTtTTgacTt|gAagtAAtAt, score=3.71
P.aeruginosa	cGCTTGAtcc|acgTCAAGCc, score=4.23
P.fluorescens	cgatTGAtct|gagTCAgggc, score=4.32

The fourth set of related sequences in this set precede various genes of the Nap complex, a periplasmic cytochrome c nitrate reductase. The arrangements are as follows:

In *P.profundum* the sequence (gCctTtATtt|ttATcActGt, score=3.94) precedes napABC, which encode periplasmic nitrate reductase precursor (EC 1.7.99.4), nitrate reductase cytochrome c550-type subunit and cytochrome c-type protein NapC.

In *V.vulnificus* and *V.parahaemolyticus* the sequence (cGtTTGAtcT|AacTCAAtCa, score=4.17; CAgTTGAcTt|gAaTCAAtTG, score=4.25) precedes *napGH*, which encode ferredoxin-type protein NapG and Polyferredoxin NapH.

In *S.oneidensis* the sequence (ACTTTGATCCCGATCGACTA, score=4.25) precedes *napDAGHB*. The gene *napD* encodes periplasmic nitrate reductase component NapD

In *C.psychrerythraea* the sequence (AaccTGATCT|AGATCAaacT, score=4.64) precedes *napEFDABC*. The gene *napE* encodes nitrate reductase component NapE protein.

Another motif exists that places the sequences preceding the *ccoN* genes, as well as other important genes, which separates the first and second sets of *ccoNOQP* in the Pseudomonads into two separate related sequence sets. The motif of this group is

(AnnnTTGATnT|AnATCAAnnnT) and the specific sequences (by the genes they precede) of the first set are:

C.psychrerythraea	*ccoN*	tttaTTGATcT|AcATCAAattt, score=4.52
I.loihiensis	*ccoN*	(none found in this set)
P.haloplanktis	*ccoN*	aAgTtTGAttg|atgTCAtAaTa, score=4.03
P.fluorescens	*ccoN*2	cagcTTGtCCg|aGGgCAAtttt, score=4.32
		CGCatTGACCc|aGGTCAtaGCG, score=3.94
P.aeruginosa	*ccoN*2	(none found in this set)
P.putida	*ccoN*2	AaccTTGcTgg|ttAtCAAtacT, score=4.22
P.syringae	*ccoN*	accatTgACGc|tCGTtAgcttg, score=4.08
	hemN	ccttTTGATcc|acATCAAggcc, score=3.82
P.profundum	*fnr*	aAtaTTGAccT|AcaTCAAatTg, score=4.77
	uspE	ActAtTGATgC|GtATCAcTttT, score=3.79
S.oneidensis	*ccoN*	cagaTTGATCT|AGATCAActtt, score=4.84
V.cholerae	*ccoN*	AAaCcTgatgg|agtgtAaGgTT, score=3.92

	fnr	aatATTGAcGt|aCaTCAATtag, score=4.24
V.parahaemolyticus	*fnr*	aataTTGATgT|AtATCAAatag, score=4.21
V.vulnificus	*fnr*	aataTTGAcgT|AtaTCAAatag, score=4.09

The second related set of sequences in this group represents the first (upstream) set of the duplicate *ccoNOQP* genes in the Pseudomonads:

P.fluorescens	*ccoN*1	cccATTGACcc|ccGTCAATact, score=4.26
P.aeruginosa	*ccoN*1	ccaATTGATCc|cGATCAATatt, score=4.84
P.putida	*ccoN*1	cccATTGAtcc|ccgTCAATacc, score=3.82

The third related set of sequences in this group precede genes encoding cytochrome d ubiquinol oxidase subunits I & II (EC 1.10.3.-) in the genomes of *C.psychrerythraea* (aAacTTGATcT|AtATCAAtaTa, score=4.47) and *V.cholerae* (AAAaTTGATac|aaATCAAaTTT, score=4.47). A gene encoding a putative transcriptional regulator arsR family is also encoded at the beginning of the operon in *C.psychrerythraea*, after the sequence.

Lastly, the fourth related set of sequences in this group precedes the *nor* genes encoding nitric oxide reductase (Nor) subunits. It occurs in three genomes:
In *P.profundum* the sequence (AAgatTGAtgc|ttgTCAtggTT, score=4.46) precedes *norCBD*, which encode Nor subunits C & B (EC 1.7.99.7) (no known metal requirements) and Nor activation protein NorD.

In *C.psychrerythraea* the sequence (aAACTTGATtt|ccATCAAGTTg, score=4.32) precedes only *norCB.*

In *P.aeruginosa* the sequence (ttgCTTGAccg|gaaTCAAGatt, score=3.88) precedes *norQE(hypothetical protein)CBD*. The genes *norQ* and *norE* encode Nor activation proteins NorQ and NorE, while the 85aa hypothetical protein encoded is homologous to a NirP in *P.stutzeri* and *P.florescens*.

The β-Proteobacteria:

Aar6 (*A.aromaticum* EbN1 817aa, gi:56478348) is one of two FUPA27s encoded in *cco* operons arranged instead in the order normally seen in those encoding FUPA29s, except that *ccoQ* is not present in FUPA29-encoding operons so that the order of transcription is:

1) *aar6* (FUPA27)

2-8) *ccoSNOQPGH*

Twenty-four bp downstream of *ccoH*, a 79aa, 2 TMS hypothetical protein is encoded codirectionally to the operon, followed 53bp later by a convergently transcribed gene encoding UspA. Upstream of the FUPA27-encoding *aar6*, 231bp away, a short-chain dehydrogenase/reductase SDR is encoded codirectionally. Divergently transcribed from the SDR gene, 199bp away, is a single gene encoding DNA polymerase, beta-like region, a 103aa protein and after that, 142bp away, the next gene is convergent and encodes hypothetical nudix hydrolase YeaB.

Interestingly, *cycZ* is also present in this genome, ~9.6kb downstream of the end of *ccoH*. It is clustered with genes encoding Coproporphyrinogen III oxidase, a Crp/Fnr family transciptional regulator, a P-type ATPase Family 5 (Cu^{+}) Transporter and a copper chaperone protein.

RegPredict analysis is not available for *A.aromaticum* at this time as the genome -and indeed the entire Order containing it (Rhodocyclales)- is not available.

Rso4 (*Ralstonia solanacearum* GMI1000 851 17545993) is the second of two FUPA27s encoded in *cco* operons arranged instead in the order normally seen in those encoding FUPA29s, except that *ccoQ* is not present in FUPA29-encoding operons so that the order of transcription is:

1) *rso4* (FUPA27 gene)

2-8) *ccoSNOQPGH*

Downstream of *ccoH*, with a 3bp overlap, an 85aa, 2 TMS hypothetical protein is encoded codirectionally to the operon, followed 21bp later by a convergently transcribed gene encoding Crp/Fnr family transcriptional regulator. Divergently transcribed from that, and thus co-directional to the cco operon, 177bp away is *cycZ*. Depending on the version of this genome used, the next gene is either 477bp away and co-directional or 207bp away but convergent; either is sufficient to assume that it is not coexpressed. Upstream of the the FUPA27-encoding *rso4*, 169bp away, there is a gene encoding putative phosphosulfolactate phosphohydrolase-related protein transcribed divergently,

followed 8bp later by a convergently transcribed gene encoding an AFG1 family ATPase.

Some 100kb away, another CcoNO pair is encoded, clustered with genes encoding copper-containing nitrite reductase (EC 1.7.2.1), cytochrome oxidase biogenesis protein Sco1/SenC/PrrC, putative copper metallochaperone and hypothetical cytochrome oxidase associated membrane protein.

In *R.solanacearum*, RegPredict suggests several possible regulatory sequences preceding the FUPA27 gene as well as related ones predicted to regulate other genes. The highest scoring one preceding the FUPA27 gene (TcCtcCTGGtt|gcCCAGtcGaA, score=6.40) is found 160bp upstream of the gene. Related sequences preceding other genes include one (tCCacCcGGCa|gGCCcGccGGg, score=4.99) preceding a two-gene operon encoding mannosyltransferase OCH1 and a 362aa putative glycosyl transferase, family 2/protein prenyltransferase. Another related sequence in this set (TCatccTcctt|gcccAttccGA, score=5.02) precedes a gene, trpD3, which encodes anthranilate phosphoribosyltransferase like protein (EC 2.4.2.18). The last member of this set (atCacccCGct|gcCGcatcGga, score=5.33) precedes a bicistron encoding D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4) and molybdopterin-guanine dinucleotide biosynthesis protein MobB.

Another set predicted by RegPredict suggests a sequence (CggGCgccGc|A|tCtctGCgtG, score=6.31) found 80bp upstream of the FUPA27 gene. Related sequences include one (cgGGtgcgcc|A|tcgctgCCgc, score=5.11) that precedes a 10-gene operon encoding:

1) aspartate aminotransferase (EC 2.6.1.1)
2) homoserine dehydrogenase (EC 1.1.1.3) (no metal requirements found for prokaryotes)
3) *thrC* encodes threonine synthase (EC 4.2.3.1)
4) uracil-dna glycosylase protein
5) *moeA*2 encodes molybdopterin biosynthesis protein MoeA
6) 6) *moaD* encodes molybdenum cofactor biosynthesis protein MoaD
7) 7) *moaE* encodes molybdenum cofactor biosynthesis protein MoaE
8) 8) *crcB* encodes a protein involved in cell processes; cell division
9) 9) D(-)-3-hydroxybutyrate oligomer hydrolase (EC 3.1.1.22) (no requirement for divalent cations [Delafield FP, et al., 1965])
10) 10) *clpB* encodes ClpB protein

The next related sequence in this set (CggGcCCgGC|A|GCgGGaCagG, score=4.96) precedes a bicistron encoding 6-hexanolactone hydrolase and a 287aa putative signal peptide protein with no significant homology. Another sequence in this set (cgcGcccgGc|G|cCgccaCcgc, score=4.85) precedes a 6-gene operon that encodes:

1) cytochrome c-type biogenesis protein CcsA/ResC
2) ATP/GTP binding protein with lipoprotein 15 domains
3) *rpoE*2 encodes a putative ECF subfamily RNA polymerase sigma-24 subunit
4) transmembrane regulator protein PrtR
5) putative sulfite oxidase subunit YedY

6) a putative 6TMS, 217aa conserved membrane protein of unknown function with a ferric reductase-like domain

Finally, a related sequence (cgCGtcGCGc|A|tCGCcgCGgc, score=5.18) precedes a single gene, *coxC*, which encodes cytochrome c oxidase polypeptide III (EC 1.9.3.1).

Nme2 (*Neisseria meningitidis* MC58 823aa, gi:15676928): is one of only three FUPA27s not encoded in some variation of a *cco* operon. Instead, it is encoded as the sixth of seven genes in an unrelated operon with another 8-gene operon encoded divergently 260bp away. The 7 proteins encoded by the operon encoding **Nme2** are, in order of transcription:

1) glutamate--cysteine ligase (EC 6.3.2.2) ($2Mg^{2+}$, $2Mn^{2+}$ or $2Cu^{2+}$ may be used, but some divalent cations are essential [Kelly BS, et al., 2002]), divergent, of Alpha- and Beta-proteobacteria type
2) 123bp later, DNA repair protein RadC
3) 73bp later, a 241aa hypothetical protein in the AANH-like superfamily with 72% homology to DNA integration/recombination/inversion protein
4) 153bp later, SCP-like extracellular family protein
5) 5bp later, GTP-binding protein HflX
6) -3bp later (overlap), Nme2 (FUPA27 P-type ATPase)
7) 162bp later, a 68aa, 2 TMS hypothetical protein with no significant homology

The next gene is transcribed convergently 393bp away. It encodes ferredoxin--NADP(+) reductase (EC 1.18.1.2) which is involved in cytochrome c oxidase biogenesis.

Divergently transcribed to this operon is another that encodes, in order of transcription:

1) 3-isopropylmalate dehydratase large subunit (EC 4.2.1.33)
2) 97bp later, putative 2-isopropylmalate synthase
3) 62bp later, 3-isopropylmalate dehydratase small subunit (EC 4.2.1.33)
4) 175bp later, DNA-cytosine methyltransferase (EC 2.1.1.37)
5) 16bp later, Type II restriction enzyme NlaIV (EC 3.1.21.4)
6) 34bp later, 3-isopropylmalate dehydrogenase (EC 1.1.1.85)
7) 163bp later, Protein yceI precursor
8) 461bp later, Aspartate ammonia-lyase (EC 4.3.1.1)

The next gene is transcribed convergently 72bp away. It encodes a 291aa, 10 TMS integral membrane protein with 57% identity to cell division protein MukB.

The genes *ccoNOP* are present in the genome and found together, although none of the gene neighborhood surrounding them is related to any genes commonly otherwise found with the cco operon. The genes *ccoGH* are also present elsewhere in the genome and found together without any other cco operon-related genes. *ccoS* is found without other cco operon genes described previously but with a gene encoding cytochrome c-type biogenesis protein DsbD, protein-disulfide reductase (EC 1.8.1.8), a relative of cycZ, encoded 2358bp away. Lastly, a homologue of the cycZ encoded in other *cco* operons is encoded in this genome as well, alone and without any *cco* operon-related genes, or anything frequently occurring for that matter, nearby.

In *Neisseria meningitidis*, RegPredict suggests several regulatory motifs preceding the beginning of the operon containing the FUPA27 gene. One sequence 31bp upstream of the gene encoding SCP-like extracellular family protein (CacCcCGtt|gcCGcGtcG, score=6.36) has one related sequence elsewhere in the genome. The related sequence (gcGCcCGtt|gcCGcGCcg, score=5.15) precedes a 5-gene operon encoding:

1) *trpE* encodes anthranilate synthase, aminase component (EC 4.1.3.27), TrpAa
2) 2) 181aa hypothetical protein in the DUF1643 superfamily
3) 3) *purK* encodes phosphoribosylaminoimidazole carboxylase ATPase subunit (EC 4.1.1.21)
4) 4) 161aa acetyltransferase, GNAT family
5) 5) *sbp* encodes sulfate and thiosulfate binding protein CysP

Another related sequence 47bp upstream of the gene encoding SCP-like extracellular family protein (TGccgTtTt|gAtAaaaCA, score=5.75) also has one related sequence elsewhere in the genome (TGtccAtTt|cAgTaagCA, score=5.56), which precedes a 3-gene operon encoding:

1) *gshA* encodes glutamate--cysteine ligase (EC 6.3.2.2), divergent, of Alpha- and Beta-proteobacteria type
2) *radC* encodes DNA repair protein RadC
3) 241aa protein with DUF208 domain in the adenine nucleotide alpha hydrolases (AANH) superfamily which is 72% identical to DNA integration / recombination / inversion protein [*Serratia odorifera* 4Rx13]

Upon examination of possible related sequence preceding the other *cco/fix* genes usually found with the FUPA27 gene *ccoI*, related sequences preceding both the operon encoding FUPA27 and *ccoNOP* were discovered. One such sequence preceding the gene encoding SCP-like extracellular family protein (cCCcgttGcCG|CGtCgggaGGa, score=6.81) by 29bp has a related sequence (aCagGCtGCgg|aaGCgGCagGa, score=6.81) 47bp upstream of this genome's *ccoN*. Additionally, another related sequence (aCggGttGccg|gaaCcgCagGa, score=5.11) precedes *accC*, which encodes biotin carboxylase of acetyl-CoA carboxylase (EC 6.3.4.14) (divalent metal ion required, Mg^{2+}, Mn^{2+} or Co^{2+} [Guchhait RB, et al., 1974]).

A second pair of sequences suggested to coregulated the *ccoI*-containing operon and that of *ccoNOP* (CCcgttGcCG|CGtCgggaGG, score=6.52; CagGCtGCgg|aaGCgGCagG, score=6.52, respectively) also has several other related sequences preceding other genes such as one (CCgtCtGaag|aagCgGcaGG, score=5.03) preceding the bicistron *ptsHI* which encodes a nitrogen regulation associated phosphocarrier protein and phosphoenolpyruvate-dependent Enzyme INtr (aka PEP-protein phosphotransferase) of the PTS system (EC 2.7.3.9) (Mg^{2+} necessary for dimerization in *E.coli*, only the dimeric enzyme can be phosphorylated; Mn^{2+} may substitute [Hoving H, et al., 1982]).

Another sequence in this set (cCggCtGccG|CatCgGatGt, score=5.03) precedes a 4-gene operon encoding:

1) homolog of *E.coli* HemY protein

2) Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
3) *hemE* which encodes uroporphyrinogen III decarboxylase (EC 4.1.1.37) (possible slight activation with Mn^{2+} [Jones RM, et al., 1993]).
4) *radA* which encodes DNA repair protein RadA

Lastly for this set is a sequence (CAAGCCGCCTAAGCGGCAGG, score=5.03) preceding a single gene encoding butyryl-CoA dehydrogenase (EC 1.3.99.2) (no known metal requirements).

A third such pair of sequences suggested to coregulate the ccoI-containing operon and that of ccoNOP (ggGTtCgGc|A|tCgGcACgg, score=6.62 and ggCTGCgGa|A|gCgGCAGga, score=6.62, respectively), though no other related sequences were found with them.

The ε-Proteobacteria

The FUPA27 of *Nitratiruptor sp.* SB155-2 (802aa, gi:152991194, accession # A6Q500) is the second of only 3 FUPA27s not encoded in some version of a *cco* operon, aside from being expressed with CcoS. Instead, it is the second gene in an operon that may be up to 17-genes long. Divergently transcribed from this operon is a 4-gene operon with the first gene of each operon overlapping by 22bp. The proteins encoded in the FUPA27-encoding putative operon, in order of transcription are:

1) asparaginase
2) -12bp later (overlap), FUPA27

3) 3bp later, Type cbb3 cytochrome oxidase biogenesis protein CcoS, involved in heme b insertion
4) 131bp later, D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase (EC 3.1.1.-)
5) 10bp later, ADP-L-glycero-D-manno-heptose-6-epimerase (EC 5.1.3.20) (no known metal requirements)
6) -3bp later, ADP-heptose synthase (EC 2.7.-.-) / D-glycero-beta-D-manno-heptose 7-phosphate kinase
7) -6bp later, Phosphoheptose isomerase 1 (EC 5.3.1.-)
8) 0bp later, glycosyl transferase, family 9
9) -7bp later, Lipid A biosynthesis lauroyl acyltransferase (EC 2.3.1.-)
10) -3bp later, a 289aa, 1TMS hypothetical protein in the Mitochondrial fission ELM 1 superfamily, its function is unknown
11) -3bp later, Glycosyltransferase
12) 164bp later, putative Glycosyltransferase
13) -7bp later, Asparagine synthetase [glutamine-hydrolyzing] (EC 6.3.5.4) (Mg^{2+} required, only Co^{2+} can replace with significant success [Boehlein SK, et al., 1994])
14) -3bp later, a 210aa hypothetical protein whose C-terminal half appears to be in the SGNH/GDSL hydrolase superfamily
15) -3bp later, ADP-heptose:LPS heptosyltransferase II

16) -16bp later, a 391aa, 12TMS hypothetical protein with 35% identity to O-antigen polymerase and a conserved domain in the Wzy-C (O-antigen ligase) superfamily
17) 400bp later, an 87aa hypothetical protein with no significant homology

Transcribed convergently to this, 219bp away, is a gene encoding a Ferredoxin--nitrite reductase (EC 1.7.7.1), the last gene itself in a 6-gene NO^{3-} and NO^{2-} ammonification operon.

There is also a divergently transcribed operon that encodes, in order of transcription:

1) 129aa, 4TMS hypothetical protein with no significant homology
2) 107bp later, Transcription termination factor Rho
3) 6bp later, Glutamate racemase (EC 5.1.1.3)
4) 23bp later, DNA polymerase III subunits gamma and tau (EC 2.7.7.7)

Convergently transcribed with a 3bp overlap is a gene encoding Endonuclease III (EC 4.2.99.18), which is the tenth gene in an operon that appears to be mostly concerned with DNA repair.

Genes encoding CcoNOQPGH are found together elsewhere in the genome with no frequently occurring proteins found adjacent to them. Two additional genes are incorporated into the operon, on either side of *ccoG*. The upstream one encodes an 87aa, 1 TMS hypothetical protein in the DUF4006 superfamily with no significant homology and the downstream one encodes a 195aa, 2TMS hypothetical protein 31% identical to 3-dehydroquinate dehydratase. Lastly, a gene encoding a DsbD/cycZ-like protein was found with no frequently occurring genes adjacent to it and only one transcribed adjacent

and codirectionally, with a 7bp overlap upstream. This gene encodes a 12TMS amino acid transporter.

The FUPA27 of *S.denitrificans* DSM 1251 (689aa, gi:78776775)) is the third of only 3 FUPA27s not encoded in some version of a *cco* operon, aside from being expressed with CcoS. Instead, it is the fourth gene of a possible 5-gene operon. The proteins encoded amongst FUPA27 are as follows, in order of transcription:

1) Phosphoheptose isomerase (EC 5.3.1.-)
2) 7bp later, D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase (EC 3.1.1.-)
3) 3bp later, asparaginase
4) 362bp later, FUPA27
5) 3bp later, Type cbb3 cytochrome oxidase biogenesis protein CcoS, involved in heme b insertion.

Upstream of this operon, a series of rRNAs are encoded which are presumed not to be expressed with this operon. Downstream of *ccoS*, 36bp away, a gene encoding a cytochrome c, class I is found.

Neither of the species described here are represented in RegPredict and thus no regulatory sequences will be discussed.

The Spirochaetes:

The FUPA27s representing the Spirochaetes, found in two species of Leptospira (the only Spirochaete genus found to contain FUPA27), are both encoded in standard *ccoNOQPGH-FUPA27-S-cycZ* operons whose only outstanding trait is that several of the cco genes are such distant homologues of those found in other phyla that SEED does not recognize their homology. A simple Psi-BLAST search reveals their relationships well however. Several hypothetical proteins surround the cco operons in these genomes; they lack significant homology to any known proteins unless otherwise stated.

Lbi9 (*Leptospira biflexa* 820aa, gi:189912906) is encoded within a standard FUPA27-encoding *cco* operon with one co-directional gene 189bp upstream encoding a 322aa, 3 TMS hypothetical protein. Divergently transcribed from that, 75bp from that is a gene encoding a 40aa hypothetical protein and 17bp away from that is a convergently transcribed gene encoding a Glutamate synthase [NADPH] small chain (EC 1.4.1.13). Downstream of *cycZ*, 18bp away, a convergently transcribed gene encodes a 108aa hypothetical protein.

Lbo1 (*L.borgpetersenii* 813aa, gi:116331093) is also encoded within a standard FUPA27-encoding *cco* operon and has a co-directional gene 651bp upstream encoding transposase IS1533 protein. Divergently transcribed from that, 37bp from that is a gene encoding a 37aa hypothetical protein and 17bp away from that is a convergently transcribed gene encoding a 60aa hypothetical protein. Downstream of *cycZ*, 84bp away, a divergently transcribed gene encodes a putative polymerase and 63bp away from that is found a convergent gene encoding a 136aa, 1TMS hypothetical protein.

No species of Spirochaetes is currently represented in RegPredict except for two of the genus Treponema, and thus, no regulatory sequences will be discussed.

The consistent and intimate association of the vast majority of FUPA27 genes with those encoding cytochrome c oxidases and related proteins, as well as genes encoding proteins involved in nitrogen fixation, leaves little doubt about the involvement of these ATPases in these processes. Their likely function is to supply copper cations to the active sites of these proteins that require them, which are consistently found embedded in the membrane. Major support for this is found in Hassani, et al. (2010) in which the putative copper translocating P-type ATPase, CtpA (TC# 3.A.3.27.4) is examined and found to be incapable of providing copper tolerance and thus is likely not involved in full export from the cytoplasm out of the cell. A similar ATPase, *fixI* in *Rhizobium meliloti* and *Bradyrhizobium japonicum* was also examined in Kahn, et al. (1989) and Preisig, et al. (1996) and found to perform a comparable function in the nitrogen fixation complex, *fixGHIS* which also encodes proteins involved in the biogenesis of heme-copper oxidases such as those involved in symbiotic nitrogen fixation. These homologous complexes are believed to be involved in the assembly of membrane (cytochrome oxidase, cbb3), periplasmic (nitrous oxide reductase, NosZ) copper enzymes. The ATPase, CtpA, appears to deliver copper to the active sites of these enzymes without effectively translocating the copper across the membrane. On the basis of these genome context analyses as well as the work reported by Hassani, et al. (2010) it is concluded that copper ATPases within the FUPA27 and FUPA29 families function to

deliver copper to enzymes, probably always integral membrane enzymes, dependent on copper for activity.

The FUPA28 ATPase Family

FUPA28 is a family of Type I P-type ATPases with 2 distinct paralogues in each species of Legionella. Those analyzed here have 847 and 852 aas and a topology of 6-7 TMS, according to TCDB. The first two or three TMS cluster around 141-226, the next two pair around 350-427 and the last 2 or 3 cluster around 681-770. The family's nearest hit in TCDB is Family 5 (TC 3.A.3.5.-), subfamilies 1 and 15 for the 852 and 847aa proteins, respectively. Through SEED all members of this family are found exclusively in Legionella strains, mapping to different regions on the chromosomes. The next closest homologues are found in related γ-Ps and are all Family 5 Cu^{+} P-type ATPases confirmed by TCDB. Conversely, no close homologues in Legionella strains are in any other family.

*RegPredict might be interesting for these proteins; however, the Order, Legionellales, is not represented in this database.

Lpn8 (*L.pneumophila... Philadelphia* 1 852aa, gi:52840486) is encoded divergently 244bp away from a single gene encoding zinc uptake regulation protein ZUR (which in turn has a convergently transcribed gene 117bp away encoding benzoylformate decarboxylate (EC 4.1.1.7) (Mg^{2+} required for activity [Mikolajek R, et al., 2007]). Downstream of the FUPA28 gene are three codirectionally transcribed genes. The first, 146bp away, encodes probable Slr0302 protein, which is a two-component response regulator with a GGDEF domain called PleD. The second, 212bp later, encodes heme oxygenase (EC 1.14.99.3) (Fe^{2+} required [Furci LM, et al., 2007]) and the third, 148bp

later, encodes an amino acid permease family protein. The next gene downstream from this is transcribed convergently and encodes a Dot/Icm (intracellular multiplication) type IV secretion system (T4SS) effector.

Its closest non-Legionella (and thus non-FUPA28) relative is a Family 5 Cu^{+} P-type ATPase (TC# 3.A.3.5.5) in *Marinobacter algicola* DG893, gi:149378220.

Lpn9 (*L.pneumophila... Philadelphia* 1 847aa, gi:52842897) is encoded divergently 360bp away from an ~11 TMS, 544aa hypothetical protein referred to as LphB [Brand BC, et al.,1994] which is thought to be a member of the TerC superfamily (TC# 9.A.30.4.1). This LphB gene is in turn met convergently 7bp away by a gene encoding IcmX protein, part of a type-IV secretion gene cluster (which is encoded in a bilateral expression unit along with IcmWV and DotA). Lpn9 is also encoded codirectionally 240bp before a gene encoding a 2 TMS, 177aa hypothetical protein with no significant homology exclusive to Legionella and is only found encoded in conjunction with the Lpn9 gene. This hypothetical protein gene is also met convergently, 73bp later by a gene encoding SnoK-like protein, which is a putative S-adenosylmethionine(SAM)-dependent methyltransferase, class I.

Its closest non-Legionella (and thus non-FUPA28) relative is a Family 5 Cu^{+} P-type ATPase (TC# 3.A.3.5.15) Synechococcus elongatus PCC 7942, gi:81300379. With such limited coexpression and no RegPredict data available, predictions for FUPA28 are restricted to cautious speculation. The proximity to a member of the FUR/ZUR family, whose members bind heavy metals such as iron and zinc and regulate transporters involved in heavy metal uptake supports the hypothesis that FUPA28s,

identified as members of the ZntA family, may actually act as Zn^{+} transporters. Association with the gene encoding PleD, which possesses a GGDEF domain, suggests that the FUPA28s' function may be related to cell surface adhesion.

Association with the genes encoding heme oxygenase and amino acid permease family proteins may possibly indicate that the FUPA28s are involved in supplying one with Fe^{2+} or working with the other as amino acid transporters, but the less frequent, though still appreciable occurrence of their genes with that of the FUPA28 make this a less feasible possibility. The small two TMS protein encoded adjacent to the second FUPA28 gene could serve as an auxiliary protein for the ATPase, as it is only encoded here, but its function is unclear as it lacks significant homology to any known proteins. The co-directional LphB gene further supports the hypothesis that FUPA28s serve a heavy metal transport, possibly resistance function. Interestingly, both FUPA28 ATPases in Legionella species appear to be encoded within gene clusters that encode constituents of ICM-type IV protein secretion systems. Whether the ATPases provide a metal ion important for secretion or prevention of phagosome-lysosome fusion, intracellular multiplication or to kill human macrophages is speculative.

The FUPA29 ATPase Family

FUPA29 P-type ATPases are generally considered to be of type I topology. They are closest to Family 5 (Copper Transporters) of the known-function P-type ATPases. Their differences however have given rise to the categorization as type VIII ATPases, along with FUPA27 [Chan, et al., 2010, p.39], because while they are of the typical type I size, many of them appear to differ slightly in topology. They are predicted by the WHAT program (but not by the TMHMM program) to have one or two additional TMS preceding TMS A [Chan et al., 2010]. They are found in some δ-Proteobacteria as well as in some members of Phylum Bacteroidetes in Classes Flavobacteria and Sphingobacteria. Only 3 of these are described in Chan, et al. (2010), one in a δ-Proteobacteria and two in Flavobacteria and only ~32 more are found in SEED, 1 in a δ-Proteobacteria closely related to the first, *Bacteriovorax marinus* SJ, and the rest in other Flavobacteria (Cellulophaga, Flavobacteria, Flavobacteriales, Gramella, Kordia, Krokinobacter, Leeuwenhoekiella, Robiginitalea, Candidatus Sulcia, Tenacibaculum/Polaribacter and Zunongwangia) or in Sphingobacteria (Cytophaga, Chitinophaga, Mucilaginibacter, Pedobacter, Algoriphagus and Sphingobacterium). *Candidatus Sulcia muelleri* SMDSEM was identified in Planctomycetes in Chan et al. (2010), but has since been classified among the Flavobacteria, although only partial genomes are represented in SEED at this time. Representatives of the 4 genera mentioned in Chan, et al. (2010) have been added here to broaden the scope of genes surrounding those encoding FUPA29.

The general topology of this family also consists of 8, 9 or (possibly) 10 TMSs, depending on if the 2 possible TMSs sometimes preceding A are absent. In any case, A & B cluster together within a 60aa region within the range of residues 167-239, TMSs 1 & 2 cluster together within a 58aa region within the range of residues 235-3033, TMSs 3 & 4 cluster about 20 residues from the center of the protein, within a 52aa region within the range of residues 420-486 and 5 & 6 cluster within the last ~60 residues, ~736-797. The "extra" TMS preceding TMS A, referred to here as A`, can be found anywhere from 13-70aa ahead of TMS A. There is very little range in size for this family, from 781-825aa overall, 797-799aa for the δ-Ps alone.

As was noted in the chapter concerning the FUPA27 ATPases, FUPAs 27 & 29 have been posited to share function based on their shared phylogeny and topology (Chan, et al., 2010) and SEED analysis also supports this. Almost all FUPA29-encoding operons START with the FUPA gene and end with a gene encoding a cycZ-like protein, as does the operon encoding the FUPA27 Rso4. This FUPA29-ccoSNOQPGH-cycZ configuration shall heretofore be referred to as the "*29-CcoSNOQPGH-cycZ* configuration" (29 snowplows?). Like the FUPA27-encoding operons, exceptions were found, however, these exceptions are merely a split in the operon following *ccoS*, usually generally not resulting in additions or removals of coexpressed genes within the two halves. This separation also brings to light a general pattern in the FUPA 27 and 29 operons for *ccoS* to follow the FUPA gene closely and for the order of the remaining operon to remain unchanged with *ccoNOQPGH* being maintained in both configurations. The other prominent difference in the operons is that *ccoN/O* is a single transcribed gene

in all FUPA29-encoding genomes. Again, the operons are highly conserved and as such will not be described entirely separately. Variations for each organism will be described using the *cco* operon as the standard and similar arrangements will be grouped accordingly.

The δ-Proteobacteria

Bba3 (*Bdellovibrio bacteriovorus* HD100 798aa, gi:42524029) is encoded within the standard operon discussed above except that *ccoH* is not present. The operon is flanked downstream by two co-directional genes and upstream by 4 divergently transcribed genes to form a greater transcription unit. The downstream genes begin 209bp downstream of *cycZ* with one encoding an OmpA/MotB domain protein which might be a flagellar motor protein which is followed 10bp later by one encoding transcriptional regulator of unknown specificity. The upstream genes begin 229bp away and encode a copper-containing nitrite reductase (EC 1.7.2.1), a probable signal peptide protein in the FGE-sulfatase superfamily 16bp later, a SCO1/SenC family protein 13bp after that and a lastly, 65bp later, an 85aa hypothetical protein in the SirA/YedF/YeeD superfamily. While it is less likely that genes that do not share a regulatory sequence are coexpressed, it is worth noting that a hydrogen peroxide inducible genes activator and an Fe-binding ferritin-like antioxidant protein/ ferroxidase (EC 1.16.3.1) (requires Fe, inhibited by Zn and terbium (Tb) [Havukainen H, et al., 2008]) are also encoded downstream after an interruption by a convergently transcribed gene encoding protein-L_isaspartate O methyltransferase (EC 2.1.1.77) and upstream, transcribed in the same

direction as the other divergently transcribed genes, is the operon described previously in the FUPA27 section, *norCBQD* followed by genes encoding methyl-accepting chemotaxis protein and a 45aa, 1TMS hypothetical protein with no significant homology. These genes are separated from the other divergent genes by two convergently transcribed genes encoding a Crp/FNR family transcriptional regulator and a 156aa hypothetical protein with no significant homology.

The only other FUPA29 found in any organism outside of Bacteroides was found in *Bacteriovorax marinus* SJ (796aa, gi: 301166690). It is given the name **Bma3** as it is a close homologue of Bba3. Like Bba3, **Bma3** is encoded within the standard *cco* operon for FUPA29s discussed above except that it is missing *ccoH*. However, different genes flank the operon on both ends with 4 downstream codirectional genes, 3 upstream co-directional genes and 5 genes transcribed divergently from the first of these. The downstream genes begin 12bp downstream of cycZ encoding a 1TMS, 39aa hypothetical protein with no significant homology, followed 37bp later by one encoding voltage-gated potassium channel beta-3 subunit, one encoding thioredoxin 3bp after that and one encoding oxidoreductase (EC 1.1.1.-) overlapping the previous one by 3bp. The 3 co-directional genes upstream encode first a Crp/Fnr family transcriptional regulator, then a 129aa hypothetical protein with homology to the CAP family effector domain (which includes such proteins as Fnr) 63bp later and finally a universal stress protein UspA 80bp after that which ends 5bp ahead of the FUPA29 gene. The divergent genes begin 218bp upstream of these, encoding first a molybdopterin biosynthesis protein MoeA, then a

581aa hypothetical protein with no significant homology 40bp later, then a 1 TMS 296aa hypothetical protein with no significant homology 249bp later, then a large hypothetical protein only found preceding the subtilisin precursor gene 151bp later and finally a subtilisin serine protease precursor (EC 3.4.21.62) (three-five Ca^{2+} binding sites, important for correct folding [Almog O, et al., 2008]). Unfortunately, unlike the greater transcription unit encoding Bba3, this one is sandwiched between numerous genes encoding short hypothetical proteins with no significant homology. As such the local gene neighborhood is not discussed.

Interestingly, Bba3 and Bma3 are more similar to their homologues in other proteobacteria, which are FUPA27s, than the FUPA29s of Bacteroides. The organisms are δ-Ps, in which FUPA27 is not found, and as the two families are posited to perform near identical, specialized functions, it seems to make sense that only one or the other be present. However, given the closer relatedness within the proteobacteria, it is suggested here that they should perhaps be grouped together, and similarities between CcoI in δ-Ps and Bacteroides be attributed to where one or the other's *cco/fix* operon originated from.

RegPredict analysis of FUPA29 in the δ-Ps is limited to *Bdellovibrio bacteriovorus* since *Bacteriovorax marinus* is not available in RegPredict at this time and no other δ-Ps have been found to encode FUPA29 proteins.

The first predicted regulatory sequence (TctTGAccgAt|tTtttTCAgaA, score=6.81) preceding the FUPA29 gene is found 149bp upstream. Several similar sequences preceding other operons in this genome were found, the first of which

(aTTTGAccGAt|tTCttTCAAAa, score=5.38) is found 188bp upstream of an operon encoding the following in order of transcription:

1. uncharacterized iron-regulated membrane protein
2. multidrug resistance protein
3. *iucA* encodes citrate:6-N-acetyl-6-N-hydroxy-L-lysine ligase, alpha subunit (EC 6.3.2.27), aerobactin biosynthesis protein IucA; siderophore synthetase large component, acetyltransferase- superfamily, group A
4. *iucB* encodes N6-hydroxylysine O-acetyltransferase (EC 2.3.1.102), aerobactin biosynthesis protein IucB; siderophore synthetase small component, acetyltransferase
5. *iucC* encodes citrate:6-N-acetyl-6-N-hydroxy-L-lysine ligase, alpha subunit (EC 6.3.2.27), aerobactin biosynthesis protein IucC; siderophore synthetase component, ligase- superfamily, group C
6. *iucD* encodes L-lysine 6-monooxygenase [NADPH] (EC 1.14.13.59), aerobactin biosynthesis protein IucD; siderophore biosynthesis protein, monooxygenase
7. TonB-dependent siderophore outer membrane receptor

The second similar sequence (GctTcAAcgt|ttatTTaAgaC, score=4.87) in this set is found 108bp upstream of an operon encoding the following genes in order of transcription:

1. lepB encodes signal peptidase I (EC 3.4.21.89)
2. pyrB encodes aspartate carbamoyltransferase (EC 2.1.3.2)
3. carA encodes carbamoyl-phosphate synthase small chain (EC 6.3.5.5)

4. 222aa hypothetical protein predicted by Glimmer/Critica with partial similarity (39% over 155 residues) to putative protein translocase
5. carB encodes carbamoyl-phosphate synthase large chain (EC 6.3.5.5)
6. phoD encodes phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1)
7. phnC encodes phosphonate ABC transporter ATP-binding protein (TC 3.A.1.9.1)
8. pilI encodes ABC-type transport system involved in multi-copper enzyme maturation permease component
9. pilD encodes leader peptidase (Prepilin peptidase) (EC 3.4.23.43) / N-methyltransferase (EC 2.1.1.-)
10. pilM encodes type IV pilus biogenesis protein PilM
11. pilN encodes type IV pilus biogenesis protein PilN
12. pilO encodes type IV pilus biogenesis protein PilO
13. pilP encodes type IV pilus biogenesis protein PilP
14. pilQ encodes type IV pilus biogenesis protein PilQ
15. 201aa hypothetical protein predicted by Glimmer/Critica with no significant homology

The third sequence in this set (aTtTTcccgAt|tTttttAAgAc, score=5.38) is found 78bp upstream of an operon encoding YjeF protein, C-terminal domain followed by a 331 hypothetical protein predicted by Glimmer/Critica with no significant homology.

The fourth sequence in this set (acTTTAccAta|atTttTAAAaa, score = 5.22) is found 45bp upstream of a single gene, *ahcY*, which encodes

adenosylhomocysteinase (EC 3.3.1.1) (no known metal requirements).

The second predicted regulatory sequence preceding the FUPA29 gene (cAgTcaTtCt|cGgAgaAaTt, score=6.68) is found 207bp upstream. Similar sequences preceding other operons in this genome include: first, one (AaaTCgatCc|cGgaaGAacT, score=5.07) 165bp upstream of an operon encoding SEC-independent protein translocase protein TATC and a putative hydrolase of the alpha/beta superfamily containing a NADPH-dependent FMN reductase superfamily domain. Second for this regulatory motif set, a related sequence (cagTcgTCCt|cGGAaaAact, score=5.07) is found 215bp upstream encoding the following genes in order of transcription:

1. *malF* encodes maltose/maltodextrin ABC transporter, permease protein MalF
2. *malG* encodes maltose/maltodextrin ABC transporter, permease protein MalG
3. *malK* encodes maltose/maltodextrin transport ATP-binding protein MalK (EC 3.6.3.19)
4. *amy* encodes alpha-amylase (EC 3.2.1.1)
5. *glk* encodes glucokinase (EC 2.7.1.2)

The third predicted regulatory sequence (TtagGAaTcg|A|aaAaTCaagA, score=6.59) preceding the FUPA29 gene is found 23bp upstream. The first related sequence (TTTAGaCTcG|A|CtAGaCTAAA, score=5.12) is found 6bp upstream of genes encoding a small-conductance mechanosensitive channel and tRNA(Cytosine32)-2-thiocytidine synthetase. The second related sequence (TatataaTaG|A|CaAaacatgA,

score = 5.12) is found 52bp upstream of a single gene encoding a HAD-superfamily hydrolase, subfamily IIB.

The fourth predicted regulatory sequence (TCTtgAccgA|T|TtttTtcAGA, score=6.59) preceding the FUPA29 gene is essentially the same as the first one described above, found 149bp upstream. The difference is that this palindrome is viewed as a 21bp sequence rather than a 22bp sequence. Both are possible lengths, and as they are not mutually exclusive, they are both discussed. The first predicted regulatory sequence (ctTttttcGG|A|CCtttgcAga, score=4.59; TaaTtAccgA|T|TtttTtAaaA, score=5.12) in this motif set is actually a pair, although the score for the first is low, found 176bp and 117bp upstream of an operon encoding the following proteins:

1. carboxylesterase type B
2. *xdhA* encodes xanthine dehydrogenase, iron-sulfur cluster and FAD-binding subunit A (1.17.1.4); xanthine oxidase (EC 1.17.3.2)
3. *xdhB* encodes xanthine dehydrogenase, molybdenum binding subunit (EC 1.17.1.4)
4. *xdhC* encodes XdhC protein (assists in molybdopterin insertion into xanthine dehydrogenase)
5. *yicP* encodes adenine deaminase (EC 3.5.4.2)
6. *phaz* encodes Poly(3-hydroxybutyrate) (PHB) depolymerase (EC 3.1.1.75)

The second predicted regulatory sequence in this motif is the same (score=5.12 here) as the first described for the 22bp-palindrome previously described, found 188bp upstream of an operon beginning with a gene encoding uncharacterized iron-regulated membrane protein and ending with one encoding TonB-dependent siderophore outer membrane receptor and thus will not be repeated exhaustively.

The third predicted regulatory sequence (TtTTGgcagA|T|TtattCAAgA, score=5.12) in this motif set is found 139bp upstream of an operon encoding the following proteins:

1. Zn-dependent protease, M50 family
2. segregation and condensation protein A
3. segregation and condensation protein B
4. 169aa hypothetical protein predicted by Glimmer/Critica with no significant homology
5. thiol-disulfide oxidoreductase-like, HTTM domain-containing protein
6. *resA* encodes a cytochrome c biogenesis (thioredoxin) maturation protein
7. 160aa hypothetical protein predicted by Glimmer/Critica with no significant homology

The fourth regulatory sequence in this motif set is another repeat (score=5.65 here) of the third mentioned for the first motif, that preceding the genes encoding YjeF and a hypothetical protein with no significant homology.

The fifth regulatory sequence (TTTtttAccA|A|TaaTttcAAA, score=5.12) in this motif is found 83bp upstream of an operon encoding the following proteins:

1. 2 TMS, 257 hypothetical protein predicted by Glimmer/Critica with no significant homology
2. histone protein
3. aminotransferase class-V family protein, putative selenocysteine lyase
4. cyclic AMP receptor protein, catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases

The sixth and final predicted regulatory sequences (gCTTcAAccg|T|ttaTTtAAGa, score=5.12; TTtTCtAcGg|T|tCaTtGAcAA, score=4.59) in this motif set are found 108bp and 53bp upstream of the first and second genes. The first of these sequences, but not the second, is a repeat of the first regulatory motif set described above, viewed in a 21-residue palindrome frame rather than the 22-residue frame, which precedes the operon containing *lebB, pyrB, phoD, phnC, carAB, pilIDMNOPQ* and two hypothetical protein genes.

A fifth predicted regulatory sequence (acCaATtTc|T|aAtATaGtg, score=6.12) is found 59bp upstream of the FUPA29 gene that has a related sequence (acaacTtTt|T|tAtAaggag, score=4.58) 48bp upstream of an operon encoding the following proteins in order of transcription:

1. Lysophospholipase (EC 3.1.1.5)
2. Alpha/beta hydrolase fold (EC 3.8.1.5)
3. 1 TMS, 178aa hypothetical protein predicted by Glimmer/Critica 54% similar over 132 residues to cytochrome c class I

4. SCO1/SenC family protein
5. *ctaC* encodes cytochrome c oxidase polypeptide II (EC 1.9.3.1)
6. *ctaD* encodes cytochrome c oxidase polypeptide I (EC 1.9.3.1)
7. *ctaE* encodes cytochrome c oxidase polypeptide III (EC 1.9.3.1)
8. 3 TMS, 111aa hypothetical protein predicted by Glimmer/Critica 56% similar over 102 residues to putative cytochrome c oxidase (caa(3)-type) subunit IV
9. *ctaB* encodes 4-hydroxybenzoate polyprenyltransferase (EC 2.5.1.-)

The sixth and last predicted regulatory sequence (GAaGtaCaT|G|AcGgcCgTC, score=6.12) for this organism is found 181bp upstream of the FUPA29 gene which has several related sequences, the first of which is (GAaGtaaaT|A|AccgcCgTC, score=5.14) 152bp upstream of an operon encoding the following proteins in order of transcription:

1. *nrfH* encodes cytochrome c nitrite reductase, small subunit NrfH
2. *nrfA* encodes nitrite reductase periplasmic cytochrome c552 precursor (EC 1.7.2.2)
3. 4 TMS, 456aa cytochrome c biogenesis protein, CcmF/CycK/CcsA family
4. 96aa hypothetical protein predicted by Glimmer/Critica with no significant homology

The next related regulatory sequences in this motif set (GAaGcgCaT|G|AaGgcCaTC, score=5.01; GAaataaTT|G|AAagcccTC, score=4.71) are found 86bp 64bp and upstream from the first and third genes, respectively, of the same

aforementioned operon containing *lebB, pyrB, phoD, phnC, carAB, pilIDMNOPQ* and two hypothetical protein genes.

Bacteroides -

Flavobacteria

Fjo5 (*Flavobacterium johnsoniae* UW101 795aa, gi:146300291) is encoded within a complete FUPA29-type *cco/fix* operon. A gene encoding coproporphyrinogen III oxidase, oxygen-independent (EC 1.3.99.22) is transcribed convergently to *cycZ* with a 3bp overlap. Upstream of this gene are several encoding hypothetical proteins, mostly with no significant homology, although one appears to be DoxX family protein. Upstream of the operon only one gene is transcribed divergently, 208bp away, encoding a Crp/Fnr family transcriptional regulator. Convergently transcribed to this, 3bp away, is a gene encoding a nucleoside triphosphate pyrophosphohydrolase MazG (EC 3.6.1.8) (no known metal ion requirements for prokaryotes).

Fps2 (*F.psychrophilum* JIP02/86 792aa, gi:150025225) is also encoded within a complete FUPA29-type *cco/fix* operon, with two notable differences. The first is that between *ccoS* and *ccoN/O*, *F.psychrophilum* has an unconserved probable acyl-ACP desaturase, stearoyl-ACP desaturase" (EC1.14.19.2) (only possible metal cofactor is Fe^{2+}, if anything [Dyer, et al., 2005]) encoded codirectionally. The second is that an 89aa hypothetical protein referred to as FIG00656051 which resembles the putative GTP-binding protein LepA is encoded codirectionally after *cycZ*, with a 12bp overlap.

Coproporphyrinogen III oxidase is encoded convergently to this gene 37bp away. Upstream, a gene encoding Crp/Fnr family transcriptional regulator is divergently transcribed alone, 109bp away. The local gene neighborhood includes a gene encoding aconitate hydratase (EC 4.2.1.3) upstream of that encoding coproporphyrinogen III oxidase and co-directional to it and several genes transcribed convergently to the gene encoding Crp/Fnr encoding such proteins as DNA mismatch repair protein MutL, two rhomboid family proteins, an endonuclease/exonuclease/ phosphatase EEP family protein and lastly a hypothetical protein which seems to be a putative RNA methylase. None of these are found to occur with the FUPA29 gene at high frequency.

Candidatus Sulcia muelleri SMDSEM encodes a FUPA29 of 806aa and gi:256370713. It is given the name **Smu2** here. The operon encoding **Smu2** is a FUPA29-type *cco/fix* operon in terms of the order of transcription, however it is unusual in that it lack *ccoQ, ccoG* and *cycZ.* The next coding region downstream of *ccoH* is 653bp away and encodes three tRNAs followed by genes encoding aspartokinase (EC 2.7.2.4) (requires Mg^{2+} or Mn^{2+} [Dungan SM, et al., 1973])/ homoserine dehydrogenase (EC 1.1.1.3), homoserine kinase (EC 2.7.1.39) and threonine synthase (EC 4.2.3.1). Due to the intergenic distance, these genes are not thought to be coexpressed. Also unusual for the operon is the presence of 3 codirectionally transcribed gene upstream of the FUPA29 gene. These genes encode LSU ribosomal protein L34p, lipoate synthase 12bp after that and GTP-binding protein EngA 1bp after that, which itself ends 14bp upstream of the FUPA29 gene. Transcribed divergently from the LSU ribosomal protein gene,

64bp away is a single gene encoding succinyl-CoA ligase [ADP-forming] alpha chain (EC 6.2.1.5) (Mg^{2+} required, the true substrate is the MgADP-complex [Joyce MA, et al., 2000]). The next coding region is 658bp away and encodes two more tRNAs. Perhaps these significant variations from the typical operon and gene neighborhood, especially concerning the placement of genes concerning translational machinery, can be explained by *S.muelleri* being an obligate "co-resident intracellular symbiont", as such is suggested by McCutheon JP, et al. (2009).

RegPredict analysis of FUPA29 in the Flavobacteria was performed using the intergenic space preceding the FUPA29 genes in both species of Flavobacterium genus, as well as the intergenic space preceding the LSU ribosomal protein L34p gene at the beginning of the FUPA29-encoding operon in *S.muelleri*, the intergenic space preceding the *ccoN/O* gene, and the intergenic space preceding the FUPA29 gene in *S.muelleri* CARI rather than that of *S.muelleri* SMDSEM, because that space is 86bp long, rather than 14bp, in the respective organisms and with respect to the predicted operon in *S.muelleri* SMDSEM beginning with the LSU ribosomal protein gene. The search was actually performed on all the of Bacteroidetes species included in this study, but only Flavobacteria members will be described in detail here. Using these as criteria, a disproportionate number of other ribosomal protein genes and tRNA genes were predicted to be co-regulated, and when the *S.muelleri*-sequences were removed, nearly all of these were eliminated from the predictions. Therefore, a combination of searches was used and results which over-predicted ribosomal protein genes and tRNA genes were disregarded. It appears that the sequence preceding the LSU ribosomal gene in

S.muelleri most certainly regulates such genes more so than *cco/fix* genes and those related to them. As such, very few strong co-regulatory predictions can be made for *S.muelleri*.

The first predicted regulatory sequence is found preceding the FUPA29 gene in both of the Flavobaterium species, 72 and 69bp upstream. A lower-scoring similar sequence preceding *ccoS* in *F.psychrophilum* is also found, 66bp upstream, as well as a similar arrangement (not described) in *C.hutchinsonii.* The sequences and scores for this motif are:

F.psychrophilum JIP02/86	*ccoI*	ttATATGAcAa|aTaTCATATtt, score=6.33
	ccoS	aAgCATGACAa|aTGTCATGtTa, score=4.59
F.johnsoniae UW101	*ccoI*	tAacATGATAa|aTATCATagTt, score=6.45

Several similar predicted regulatory sequences preceding other genes/operons are found. The first such sequence for this motif occurs in *F.psychrophilum* as well as *Algoriphagus sp.* PR1. The same operon is preceded in both species. In *F.psychrophilum*, the sequence (AaAaATGATtA|TtATCATaTaT, score=5.09) precedes the *feoAB* genes, which encode ferrous iron transport proteins FeoAB, by 72bp.

The second such sequence for this motif occurs only in *F.johnsoniae*. The sequence (AAAagTaATAA|ATATcAtaTTT, score = 5.51) is found 94bp upstream of the genes *hemHJ*, which encode protoporphyrin ferrochelatase, protoheme ferro-lyase, final enzyme of heme biosynthesis (EC 4.99.1.1) (contains a [2Fe-2S] cluster [Shepherd M, et al., 2006]) and protoporphyrinogen IX oxidase, novel form, HemJ (EC 1.3.-.-), respectively.

The third such sequence for this motif also occurs only in *F.johnsoniae*. The sequence (tttTATGATAg|tTATCATAttt, score=5.48) is found 68bp upstream of a 5-gene operon encoding the following proteins:

1. *fabH1* encodes beta-ketoacyl/3-oxoacyl-acyl-carrier-protein synthase I, KASIII (EC 2.3.1.41) (stabilized by Mg^{2+} [Price AC, et al., 2003])
2. 175aa hypothetical protein containing a methyltransferase domain which has 83% similarity to RNA polymerase sigma-70 factor [*Bizionia argentinensis* JUB59], Length=175aa
3. glycosyl transferase, family 2
4. ABC-type transport system permease component, possibly involved in resistance to organic solvent, (TC 3.A.1.27: the γ-Hexachlorocyclohexane (HCH) Family)
5. ABC-type transport system permease component, ATPase component, possibly involved in resistance to organic solvent, (TC 3.A.1.27)

The fourth such sequence for this motif also occurs only in *F.johnsoniae*. The sequence (tttaTtGAaaa|aaaTCtAattt, score=5.25) is found 48bp upstream of an 8-gene operon encoding the following proteins:

1. *cox15/cyoE* encodes heme O synthase, protoheme IX farnesyltransferase (cytochrome oxidase assembly factor) (EC 2.5.1.-) COX10-CtaB
2. alternative cytochrome c oxidase polypeptide CoxO (EC 1.9.3.1); Cytochrome c oxidase, subunit III (EC 1.9.3.1)
3. *cox3* encodes alternative cytochrome c oxidase polypeptide CoxP (EC 1.9.3.1); Cytochrome c oxidase, subunit III (EC 1.9.3.1)

4. 3 TMS, 116 aa hypothetical protein with Prokaryotic Cytochrome C oxidase subunit IV (COX4_pro) superfamily domain
5. 1 TMS, 218aa hypothetical protein with 37% similarity to protein SCO1/SenC
6. cytochrome oxidase biogenesis/electron transport protein Sco1/SenC/PrrC, putative copper metallochaperone
7. *yozB* encodes YozB protein (involved in respiratory electron transport chain)
8. 2 TMS, 90aa hypothetical protein with no significant homology

The next predicted regulatory sequence is suggested to regulate the FUPA29-encoding operons found in *F.psychrophilum, F.johnsoniae, P.heparinus* and *C.hutchinsonii*. The Flavobacteria sequences and distances are found below:

F.psychrophilum		70bp upstream of *ccoI*	ATATGAcAa|aTaTCATAT, score=6.41
		64bp upstream of *ccoS*	gCATGACAa|aTGTCATGt, score=4.84
	245bp	(EC1.14.19.2) gene	gCATGACAa|aTGTCATGt, score=4.84
	62bp upstream of	*hemN*	AcATGATAA|TTATCATtT, score=5.55
F.johnsoniae		67bp upstream of *ccoI*	acATGATAa|aTATCATag, score=6.41

Several similar predicted regulatory sequences preceding other genes/operons are found. The first three such sequences for this motif are repeats of the first three operons of the previous set and thus will not be described again in detail.

The first novel member of this motif set is found preceding operons in both *F.psychrophilum* and *S.muelleri.* In *F.psychrophilum,* the sequence (ATATaATAa|aTATaATAT, score=5.38) precedes three genes by 18bp (and thus includes the TATA box), which encode the following proteins:

1. *rpbA* encodes RNA-binding protein

2. *dnaE/dnaQ* encodes DNA polymerase III alpha subunit (EC 2.7.7.7) (Mg^{2+} required [Tuske S, et al., 2000]).
3. Microcin C7 self-immunity protein mccF

In *Candidatus S.muelleri*, a similar sequence (aaAagATAT|ATATaaTac, score=4.76) precedes a larger but similar operon by 51bp which encodes the following proteins:

1. *ackA* encodes acetate kinase
2. *trmE* encodes tRNA modification GTPase TrmE
3. putative bifunctional DNA polymerase III, alpha subunit (DnaE)/DNA polymerase III, epsilon subunit (DnaQ)
4. putative bifunctional argininosuccinate synthase (ArgG)/acetyltransferase activity of N-acetylglutamate synthase (ArgA)
5. *argC* encodes N-acetyl-gamma-glutamyl-phosphate reductase
6. *argD* encodes acetylornithine aminotransferase
7. *carA* encodes carbamoyl phosphate synthase small subunit
8. *carB* encodes carbamoyl phosphate synthase large subunit
9. *argF* encodes putative ornithine carbamoyltransferase
10. *argB* encodes acetylglutamate kinase
11. putative acetylornithine deacetylase (ArgE)/succinyl-diaminopimelate desuccinylase (DapE)

The next member of this set is found in *F.johnsoniae* only. The sequence (aAATGATaa|aaATCATTg, score=5.33) precedes a single gene, *trxB2*, by 57bp which

encodes thioredoxin reductase (EC 1.8.1.9) (contains redox active cluster $[Fe_2S_2]^{2+/+}$ [Walters EM, et al., 2005])

The next member of this set is found only in *Candidatus S.muelleri*. The sequence (AaATaATAT|ATATaATaT, score=5.25) precedes a bicistron by 131bp which encodes thioredoxin-disulfide reductase and putative tRNA:rRNA methyltransferase, whose corresponding gene is *trmH*.

The next member of this set is found in *F.psychrophilum* and precedes a very long operon. The sequence (AaaTgACAa|aTATaAaaT, score=5.25) is found 131bp upstream of the operon, which encodes the following proteins in order of transcription:

1. 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)
2. *fabB* encodes 3-oxoacyl-[acyl-carrier-protein] synthase, KASI (EC 2.3.1.41)
3. Acyl carrier protein
4. Predicted acyltransferase
5. Phytoene desaturase, neurosporene or lycopene producing (EC 1.3.-.-)
6. *darA* encodes Dialkylrecorsinol condensing enzyme
7. *darB* encodes 3-oxoacyl-[acyl-carrier-protein] synthase III (EC 2.3.1.41)
8. *darC*1 encodes (3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase (EC 4.2.1.-)
9. *darC*2 encodes Probable acyl carrier protein involved in flexirubin-type pigment biosynthesis DarC2
10. 134aa hypothetical protein with no significant homology

11. 332aa hypothetical protein with 45% similarity to lantibiotic protection ABC superfamily ATP binding cassette transporter [*Eubacterium yurii subsp. margaretiae* ATCC 43715], Length=327
12. *hpaIIM* encodes Modification methylase HpaII (EC 2.1.1.37) (no known cation requirements for prokaryotes)
13. *hpaIIR* encodes Type II restriction enzyme HpaII (EC 3.1.21.4) (Mg^{2+} required [Orlowski J, et al., 2008])
14. Export ABC transporter ATP-binding protein
15. ABC-type multidrug transport system, permease component
16. 4-hydroxybenzoyl-CoA thioesterase family active site
17. 3-oxoacyl-[acyl-carrier-protein] synthase, KASII (EC 2.3.1.41)
18. 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) (binds NADP+, does not use metal cofactor [Silva RG, et al., 2006 & 2008])
19. Acyl carrier protein
20. S23 ribosomal protein
21. 3-oxoacyl-[acyl-carrier-protein] synthase, KASII (EC 2.3.1.41)
22. 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)
23. Polysaccharide deacetylase family protein
24. LolA-like outer-membrane lipoprotein carrier protein
25. 182aa hypothetical protein with no significant homology
26. 1 TMS, 208aa hypothetical protein 42% similar to outer membrane insertion C-signal domain protein

27. (3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase (EC 4.2.1.-)
28. 161aa hypothetical protein with no significant homology
29. Glycosyl transferase, group 2 family protein
30. membrane protein, inferred for ABFAE pathway
31. Carotenoid cis-trans isomerase (EC 5.2.-.-)
32. peptidase C45, acyl-coenzyme A:6-aminopenicillanic acid acyl-transferase (EC 2.3.1.164) (no known metal requirements)
33. Phenylacetate-coenzyme A ligase (EC 6.2.1.30) (Mg^{2+} required, Mn^{2+} may substitute [Erb TJ, et al., 2008]); Coenzyme F390 synthetase
34. tRNA dihydrouridine synthase B (EC 1.-.-.-)

The next member of this set is found in *F.johnsoniae* and precedes an operon that includes several unique cytochrome-related genes. The sequence (AAATGACAa|gTGTCATTT, score=5.20) gene:

1. Cytochrome d ubiquinol oxidase subunit I (EC 1.10.3.-)
2. Cytochrome d ubiquinol oxidase subunit II (EC 1.10.3.-)
3. DNA double-strand break repair rad50 ATPase
4. GepA protein
5. 2 TMS, 145aa hypothetical protein with no significant homology
6. plasmid maintenance system antidote transcriptional regulator protein, XRE family
7. Acetyltransferase, GNAT family

The next predicted regulatory sequences (aTTGTtttac|A|aatttACAAa, score=5.37; tTTGaTgcAa|A|aTtaAcCAAt, score=5.77) preceding FUPA29-encoding operons in this group are found in *F.psychrophilum* and *F.johnsoniae,* 94bp and 192bp upstream of the *ccoI* gene, respectively.

Several related sequences are found elsewhere in these genomes which are predicted to coregulate the genes they precede. The first predicted regulatory sequence (AtTgtTgtaa|A|aattAcgAtT, score=5.57) for this group is found in *F.johnsoniae,* 117bp upstream of a bicistron encoding TonB-dependent receptor and putative G-D-S-L family lipolytic protein.

The next member of this motif set (tttttTgtaa|A|aataAtcttt, score=5.49) is found in *F.psychrophilum,* 62bp upstream of a bicistron encoding Magnesium and cobalt transport protein CorA and Acid-resistant locus arl7 (Fragment).

This concludes the consistently predicted coregulated genes. Several other slightly lower scoring sequences were also found, but the generally reinforce the same predictions and thus will not be exhaustively described. As noted above, *S.muelleri* has nearly no relevant predictions, most likely because the available sequence is actually regulatory sequence concerning the ribosomal subunit protein gene closely preceding the *cco* operon. Whether this is because the *cco* operon serves a ribosome-related purpose in this organism or not may be speculated on, but considering that this relationship is found nowhere else in the organisms available to survey in SEED, it is assumed to be a coincidence unique to this organism. No associated functional indications, other than

perhaps a reduced importance to tight regulation of the *cco* operon in general will be made here.

Sphingobacteria

Cytophaga hutchinsonii ATCC 33406 encodes a FUPA29 of 799aa and gi:110637541. It is given the name **Chu2** here. The operon encoding **Chu2** is a FUPA29-type *cco/fix* operon in terms of the order of transcription, however it is unusual in that it is interrupted by two co-directional genes and one counter-directional one, all between the FUPA29 gene and *ccoS*. The two co-directional genes begin 153bp downstream of the FUPA29 gene with one encoding a 92aa hypothetical protein with no significant homology which is followed 190bp later by one encoding a 264aa universal stress protein A. Convergently transcribed to this gene, 101bp away, a gene encoding delta-9 fatty acid/stearoyl-CoA 9 desaturase (EC 1.14.19.1) (no special ion requirement [Foot M, et al., 1983]) is found, itself transcribed divergently from ccoS, 283bp away. The rest of the operon is unchanged. Following *cycZ*, 4 genes are found transcribed convergently, only the first of which is found near the operon at high frequency, 6bp away, and encoding coproporphyrinogen III oxidase. The other three encode probable RND family efflux transporter MFP subunit 170bp later, RND multidrug efflux transporter/acriflavin resistance protein 12bp after that and lastly a 2 TMS, 162aa hypothetical protein with no significant homology 358bp after that. Two genes are also transcribed codirectionally upstream of the FUPA29 gene, and 5 genes are transcribed divergently from those. The co-directional upstream genes encode PAS/PAC sensor

signal transduction histidine kinase and Crp/Fnr family transcriptional regulator, with a 15bp space between the genes and a 29bp space before the FUPA29 gene; both are found near the operon at high frequency. The divergently transcribed genes encode phenylalanyl-tRNA synthetase beta chain (EC 6.1.1.20) (requires Mg^{2+} [Barrett AR, et al., 2008]) 317bp away, a 99aa hypothetical protein referred to as FIG00412596 with no significant homology 98bp later, putative cell division protein ZapA 22bp after that, 2',3'-cyclic-nucleotide 2'-phosphodiesterase (EC 3.1.4.16) (enzymatic activity is stimulated by Co^{2+} and Mn^{2+} but not stimulated by Ca^{2+}, Mg^{2+}, Fe^{2+}, Fe^{3+}, and Zn^{2+} [Kimura Y, et al., 2009]) 73bp after that and finally a 1 TMS, 69aa hypothetical protein referred to as FIG00574234 with no significant homology overlapping the previous gene by 10bp. None of these genes are found at significant frequency near the operon.

Algoriphagus sp. PR1 encodes a FUPA29 of 808aa and gi:311746216. It is given the name **Asp2** here. The operon encoding **Asp2** is also a FUPA29-type *cco/fix* operon and it is continuously transcribed intact. The *cycZ* gene is transcribed convergently with a coproporphyrinogen III oxidase encoding gene, with a 3bp overlap. Three proteins are encoded codirectionally upstream of the coproporphyrinogen III oxidase encoding gene. They are a putative von Willebrand factor type A vWFA, a 1 TMS, 298aa hypothetical protein with no significant homology and bacteroides aerotolerance protein BatA. Two genes are found codirectionally transcribed upstream of the FUPA29 gene, encoding hypothetical proteins of 39 and 208 residues in length, the second gene overlapping the first by 64bp and 67bp away from the FUPA29 gene. Upstream of these are encoded

large and small ribosomal subunits and a 5S RNA and past those is a 39aa hypothetical protein with no significant homology. None of these are found at significantly high frequencies to suggest meaningful co-regulation.

The Genus Pedobacter includes 3 species found in SEED: *P.saltans* DSM 12145, *P.heparinus* DSM 2366 and *P.sp.* BAL39. However, in the latter two species, the *cco/fix* operons and their genetic neighborhoods are nearly identical, and thus shall be treated as one with inclusion of the minor differences.

P.saltans DSM 12145 encodes a FUPA29 of 794aa and gi:325103488. It is given the name **Psa2** here. The operon encoding **Psa2** is a small fragment of the standard FUPA29-type *cco/fix* operon, encoding only the FUPA29 and CcoS. One small additional gene is transcribed codirectionally downstream and two small genes are transcribed divergently from the FUPA29 gene. The downstream gene encodes a putative histone H1-like protein Hc1 161bp downstream of *ccoS* and the two genes divergently transcribed begin 326bp away with the first encoding a 57aa hypothetical protein with no significant homology and the second, 406bp further, encoding a 39aa hypothetical protein with no significant homology. Convergently encoded to these, the same large and small ribosomal subunits, 5S RNA and 41aa hypothetical protein, all homologous to those found in *Algoriphagus sp.* PR1 above, although there the hypothetical protein is only 39aa and the the homology of the two small proteins is tenuous with 40% identity and 57% positive homology. It is interesting, though not indicative of related function, that these two species share some of the same unusual

upstream gene neighborhood characteristics around these FUPA29 genes. However whereas the rest of the *cco/fix* operon is still found immediately downstream in *A.sp*. PR1, instead it is found 470kb away in *P.saltans*.

The rest of the operon is found intact as *ccoN/OQPGH-cycZ,* with *cycZ* overlapping by 29bp with a convergently transcribed gene encoding a 46aa hypothetical protein with no significant homology. The hypothetical protein is in turn transcribed divergently 167bp from another single gene, which is thus transcribed codirectionally to *cycZ,* 278bp away from it. This gene is of some interest because it encodes a 402aa hypothetical protein that seems to be partially homologous to shorter proteins (150-200aa) such as a copper metallochaperone bacterial analog of the Cox17 protein for its N-terminal portion as well as cytochrome oxidase biogenesis protein Sco1/SenC/PrrC a putative copper metallochaperone for its C-terminal portion. Convergently transcribed to this gene, 189bp away, is one encoding the perennial coproporphyrinogen III oxidase that is preceded codirectionally by three other genes. These genes encode a putative PAS/PAC sensor signal transduction histidine kinase with similarity to fixL (32% identity, 48% positive homology to gi:32473302), the also perennial Cpr/Fnr family transcriptional regulator and D-lactate dehydrogenase (EC 1.1.1.28) (Ca^{2+} or Mg^{2+} required [Erwin AL, et al., 1993]). Upstream of ccoN/O, there are 5 genes that may compose a single divergent operon starting 179bp away. The genes encode the following proteins in order of transcription:

1) 1 TMS, 108aa hypothetical protein
2) 134bp later, sulfate permease

3) 21bp later, carbonic anhydrase (EC 4.2.1.1) (zinc metalloenzyme, Co^{2+} can substitute for Zn^{2+} [Elleby B, et al., 2001])
4) 142bp later, a 3 TMS, 177aa hypothetical protein with no significant homology
5) 252bp later, cytosolic long-chain acyl-CoA thioester hydrolase family protein.

Another operon meets this one convergently after this gene, but as their frequency of occurrence near the operon is low, they will not be discussed.

As mentioned above *P.heparinus* DSM 2366 and *P.sp.* BAL39 encode nearly identical *cco/fix* operons and surrounding gene neighborhoods. Their FUPA29s are given the names **Phe2** (799aa, gi:255532376) and **Psp2** (800aa, gi:149277721) here. The entire *29-ccoN/OQPGH-cycZ* operon is intact; with a coproporphyrinogen III oxidase encoded convergently 38 or 75bp away. The coproporphyrinogen III oxidase gene is preceded by the same 3 genes here as in *P.saltans*, PAS/PAC sensor signal transduction histidine kinase with similarity to fixL (31% identity, 50% positive homology to gi:32473302), Cpr/Fnr family transcriptional regulator and D-lactate dehydrogenase. Transcribed divergently from these another gene common to all the Pedobacter *cco/fix* gene neighborhoods is found transcribed alone encoding a 591-600aa hypothetical protein referred to as FIG00906767 but with no significant homology to suggest function. Divergently transcribed from the operon, 164 to 232bp away, 5 genes are found transcribed in common in both species. The proteins they encode in the order of transcription are:

1) metal-dependent phosphohydrolase, HD subdomain

2) 6 or 71bp later, polyphosphate kinase (EC 2.7.4.1) (Mg^{2+} required, Mn^{2+} and Zn^{2+} can substitute with 22% and 10% effectiveness respectively, Ca^{2+} cannot [Ahn K, et al., 1990])
3) -3 or 12bp later, 260-293aa hypothetical protein: FIG00908596/SdiA-regulated family protein
4) 36-116bp later, 784-786aa hypothetical protein: FIG00906934
5) 7-14bp later, surface antigen (D15) precursor

Each species also has one additional gene in these operons that is unique to them. In *P.heparinus*, a 39aa hypothetical protein is encoded at the beginning of the operon, overlapping the FUPA29 gene by 44bp and *P.sp* BAL39 has manganese superoxide dismutase (EC 1.15.1.1) (may bind $1Cu^{+}$ and $1Zn^{+}$ each [Beyer W, et al., 1991]) encoded 324bp after the end of the operon.

P.heparinus also has a thioesterase-like protein encoded convergently to the surface antigen gene and two more genes transcribed divergently from that which encode a twin-arginine translocation pathway signal protein and a glucose-methanol-choline (GMC) oxidoreductase:NAD binding site. While the *P.sp* BAL39 region lacks these 3 genes, there is an operon transcribed in the same direction as the *cco/fix* operon common to both species found past these three genes in *P.heparinus* and instead of them in *P.sp* BAL39. The operon is the *thiSCEDEGHF* operon with the addition of a gene encoding RNA binding S1 domain-containing protein at the end. This operon occurs at sufficiently low frequency near the cco/fix operon to assume that their expression is not intimately tied, but it is still found frequently enough to be worth noting.

RegPredict analysis of FUPA29 coregulation in the Sphingobacteria was performed using the intergenic sequence preceding the *cco* operon. In the case of *C.hutchinsonii*, this sequence was extracted both from upstream of the PAS/PAC sensor gene (*barA*) and *ccoS*; in *P.saltans*, this sequence was extracted from upstream of *ccoI* and *ccoN/O*. In all other cases the sequence was extracted from directly upstream of the *ccoI* gene only. Also, *P.saltans* is not available in RegPredict and thus, although the sequence is used to expand the possible search set, no RegPredict analysis of the genome was performed.

The first predicted regulatory sequence preceding FUPA29-encoding operons was found in three species, with distance upstream and scores as follows:

P.sp. BAL39	215bp	*ccoI*	tatttCctTAC|GTAtaGgtcat, score=4.56
C.hutchinsonii	91bp	*barA*	ttcgTCataTC|GAatgGAtttt, score=5.10
A.sp. PR1	126bp	*ccoI*	tttAGCtTTtA|TgAAgGCTctc, score=4.58

Several similar predicted regulatory sequences precede other operons/genes throughout these genomes. The first such related sequence (cAtttCgtTtc|acAtgGcttTt, score = 4.87) is found in *Pedobacter sp.* BAL39, 74bp upstream of an operon encoding the following proteins:

1. putative Fe^{2+}-dicitrate sensor FecR, anti-FecI sigma factor
2. 1 TMS, 412aa hypothetical protein with no significant homology
3. TonB-dependent outer membrane receptor
4. 402aa hypothetical protein, DUF4374 superfamily

5. putative iron-regulated membrane protein with PiuB domain, 55% similar to TonB-dependent receptor plug [*Mucilaginibacter paludis* DSM 18603] over 366 residues
6. NADH dehydrogenase (quinone)
7. Na^+:H^+ antiporter MnhB subunit-related protein
8. NADH-ubiquinone oxidoreductase, chain 4L
9. NADH dehydrogenase (quinone)
10. Na^+:H^+ ion antiporter family protein
11. monovalent cation/proton antiporter, MrpF/PhaF family protein
12. monovalent cation/proton antiporter, MnhG/PhaG subunit

The next predicted regulatory sequence (ttctTCgtTTa|aAAttGAtttc, score=5.18) in this motif set is found in *P.heparinus* DSM 2366 and precedes a tricistron, by 222bp, that encodes Cold-shock protein DNA-binding, NAD-dependent epimerase/dehydratase and beta-lactamase domain protein

The second predicted regulatory sequences preceding FUPA29-encoding operons in this group are found in three species, with distance upstream and scores as follows:

P.heparinus	164bp	*ccoI*	tTGAAagccgG|CatcggTTCAt, score=5.00
C.hutchinsonii	144bp	*ccoS*	gataaATaCtG|CtGaATgccat, score=4.56
	190bp	*barA*	AtGaAatatgG|CtgccgTgCtT, score=5.47
A.sp. PR1	98bp	*ccoI*	AtGatAAcCtt|tcGtTTtaCtT, score=4.43

Several related regulatory sequences are found preceding other operons/genes in

these genomes. The first such sequence (AtGatagatgG|CttaggtgCtT, score=4.86) is found in *Algoriphagus sp*. PR1, 41bp upstream of an operon encoding the following proteins:

1. 13 TMS, 467aa hypothetical protein, possible MFS permease
2. mannosyl transferase
3. Probable SAM-dependent methyltransferase
4. Glycosyltransferase
5. Glycosyltransferase-like protein
6. *osmC* encodes OsmC family, osmotically inducible protein

The next predicted regulatory sequence (AAGAAatACgg|taGTtgTTCTT, score=5.06) in this set is found in *C.hutchinsonii*, 205bp upstream of a bicistron encoding Ferrichrome-iron receptor and a 68aa hypothetical protein with no significant homology.

The next predicted regulatory sequence (cAtaAAtctgg|ttgtgTTccTt, score=4.87) in this set is found in *Pedobacter sp*. BAL39, 136bp upstream of a single gene encoding transcriptional regulator, TetR family protein.

The third predicted regulatory sequences preceding FUPA29-encoding operons in this group are found in two species in this group, as well as *F.psychrophilum*, 72bp upstream of *ccoI*. The distances upstream and scores are as follows:

P.sp. BAL39	173bp	*ccoI*	tAaCaGtATAa|aTATgCgGgTt, score=4.40
C.hutchinsonii	229bp	*ccoS*	tCATGataaAa|aTcataCATGc, score=5.01

The first related regulatory sequence (gAATcgtaaAa|aTaatacATTt, score=4.91) in this motif set is found in *C.hutchinsonii*, 69bp upstream of a bicistron encoding

putative sugar transferase with Ferritin-like domain and 267aa hypothetical protein with Short C-terminal domain.

The next member of this set is a sequence (tgtTgataaAa|aTaatatAttc, score=5.04) found in *Pedobacter sp*. BAL39 which precedes a monocistron, by 51bp, that encodes C4-dicarboxylate transport transcriptional regulatory protein.

The next member of this set is a sequence (GAAaaatAcAa|aTaTtaaaTTC, score=5.19) found in *C.hutchinsonii* which precedes a monocistron called *malZ*, by 101bp, that encodes Alpha-glucosidase, family 31 of glycosyl hydrolases, COG1501.

The fourth predicted regulatory sequences preceding FUPA29-encoding operons in this group are found in three species, with distance upstream and scores as follows:

P.heparinus	71bp	*ccoI*	ATCagttcC|T|GaccagGAT, score=4.76
P.sp. BAL39	78bp	*ccoI*	tTCTatcaG|C|CaagaAGAt, score=4.40
	65bp	PAS/PAC gene	tTtcgatgg|T|gatcaagAt, score=4.67
C.hutchinsonii	72bp	*ccoS*	ATGcatTgG|T|CaAtctCAT, score=4.41
	127bp	*barA*	tTtGTgtaC|T|GataACgAt, score=4.58

Several related sequences are found preceding other operons/genes in the genomes of interest. The first members of this motif set are also found in three genomes. The sequences precede two identical operons in the Pedobacter species and a related operon in *A.sp.* PR1. The Pedobacter operons are described as follows:

P.heparinus DSM 2366	250bp	*dfrA*	ATGccttaC|T|GaccatGAT, score=4.41
Pedobacter sp. BAL39	250bp	*dfrA*	ATGcattgc|T|caccatGAT, score=4.85

1. *dfrA* encodes Dihydrofolate reductase (EC 1.5.1.3) (monovalent cations activate [Nixon PF, et al., 1968])

2. *secDF* encodes Protein-export membrane protein SecD (TC 3.A.5.1.1) / Protein-export membrane protein SecF (TC 3.A.5.1.1)
3. *ndh* encodes NADH dehydrogenase (EC 1.6.99.3) (no known metal cofactors)
4. *leuA/mvaB* encodes Hydroxymethylglutaryl-CoA lyase (EC 4.1.3.4) (Mn^{2+} required, Mg^{2+} can substitute [Forouhar F, et al., 2005])

In *Algoriphagus sp.* PR1, the sequence (ATCtttttG|C|CttcagGAT, score=4.23) is found 157bp upstream of a large, partially repetitive operon which encodes the following proteins:

1. *udk* encodes Uridine kinase (EC 2.7.1.48) (requires Mg^{2+} [Orengo A, et al., 1978])
2. 103aa hypothetical protein with no significant homology
3. secDF encodes Protein-export membrane protein SecD (TC 3.A.5.1.1) / Protein-export membrane protein SecF (TC 3.A.5.1.1)

The next member of this motif set is found in *P.heparinus.* The sequence (tttCTttTC|T|GAttAGggt, score=4.66) is found 173bp upstream of an operon encoding the following proteins:

1. alkylhydroperoxidase like protein, AhpD family
2. YCII (conserved motif)-related protein
3. RNA polymerase sigma factor, sigma-70 family
4. Glutamate dehydrogenase (NADP(+))
5. L-lysine 6-transaminase

6. 1 TMS, 134aa hypothetical protein 53% similar to putative Glu/Leu/Phe/Val dehydrogenase family protein over 128 residues

The next member of this motif set is found in *C.hutchinsonii*. The sequence (ATtggttgg|T|gatgatgAT, score = 4.71) is found 226bp upstream of a bicistron encoding Protoporphyrinogen IX oxidase, aerobic (EC 1.3.3.4) and Dolichol-phosphate mannosyltransferase, encoded by genes *ppo* and *lgtD*, respectively.

The fifth predicted regulatory sequences preceding FUPA29-encoding operons in this group are found in three species, with distance upstream and scores as follows:

P.heparinus	164bp	*ccoI*	tTGAAagccgG|CatcggTTCAt, score=5.09
C.hutchinsonii	190bp	*barA*	AtGaaatatgG|CtgccgtgCtT, score=5.28
	223bp	*ccoQ*	AaaaAaGataG|CagcCaTgcgT, score=4.69
A.sp. PR1	120bp	*ccoI*	ttttAtGAagG|CtcTCtTtggg, score=4.79

The first predicted regulatory sequence in this motif set is in *C.hutchinsonii* as well as *F.johnsoniae.* It is a repeat of the very long operon described above in the Flavobacteria RegPredict section and thus will not be repeated. The sequence (AAatAacatgG|CagcagTtcTT, score=4.55) precedes the Esterase (EC 3.-.-.-) encoding gene of the operon by 34bp. The operon encodes other proteins including two copies each of 3-oxoacyl-[acyl-carrier-protein] synthase, KASII (EC 2.3.1.41) and 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) as well as Acyl carrier protein, Polysaccharide deacetylase, LolA-like outer-membrane lipoprotein carrier protein, Glycosyl transferase, group 2 family protein and the membrane protein inferred for ABFAE pathway.

The suggestion that several of the above mentioned genes predicted for co-regulation through RegPredict are in fact reinforced by their recurrence using unique sequence sets to search. While we find an abundance of additional cytochrome-related proteins are likely to be coexpressed with the *cco/fix* genes, including auxiliary/alternative copies, the even more prevalent pattern observed is the recurrence of proteins dealing with iron-handling and usage. Not only do several ferredoxin-type proteins come to light, but iron concentration regulating genes such as FecR, which detects Fe^{2+}-citrate and shuts down the respective transporter, FecI, as well as various TonB-related charged siderophore importers are found. Taken out of context, one could mistakenly assume these to be indications of related Fe^{2+} function in FUPA29. However, the presence of the FUPA29 gene within the *cco/fix* operon, and consideration of the cation demands not only of Cco-[$2Cu^{+}$] but of the greater respiration apparatus involved, including cytochrome c itself which has a heme-[Fe^{2+}] center, supports the hypothesis that the putative array of coregulated genes described work in concert, perhaps to restrict Cco translation to times when the availability of *both* copper and iron are such that electron transport can function properly.

The FUPA30 ATPase Family

FUPA30 is a family of Type II P-type ATPases with 4 subfamilies representing (in order) orthologues in Proteobacteria, Bacteroidetes, Spirochaetes and Cyanobacteria. Those analyzed here range in size from 825-867 aas and have topologies of 9-10 TMS according to TCDB, although those generally characterized as having 9 appear to also have 10 upon closer inspection. The first two, which correspond to the standard TMS 1 & 2 notation and which may sometimes be mistaken for one, cluster within residues 49-89, 3 & 4 cluster within 225-299, 5 & 6 cluster within 641-704, 7 & 8 cluster within 701-782 and 9 & 10 cluster within 764-852. The family's nearest hit in TCDB is Family 2 (TC 3.A.3.2.-), subfamilies 21 and 28 being the closest. It is for this reason that it was suggested in Chan, et al. (2010) that FUPA30 proteins might transport Ca^{2+}. Also noted in Chan, et al. (2010), the FUPA30 proteins are ~50 residues shorter than the average Type II P-type ATPase, apparently due to an abrogation of an N-terminal segment normally present in Type II enzymes.

TC# 3.A.3.30.1 contains 4 proteins found in α-, β- and δ-proteobacteria ranging in size from 825-896 aas and represented by the 4 proteins below.

Bja1 (*Bradyrhizobium japonicum* USDA 110 850aa, gi:27378926) is encoded by a gene not found to be significantly co-localized with any other genes. There is one gene 442bp upstream, codirectionally transcribed that encodes a Crp/Fnr family transcriptional regulator. Divergently transcribed from this, 30bp away, is a gene encoding a 67aa hypothetical protein with no significant homology. Downstream of the

FUPA30 gene, 3 genes are found transcribed convergently to the FUPA30 gene, the last one overlapping the FUPA30 gene by 3bp. These genes encode, in order of transcription, a bipolar DNA helicase, an alkylhydroperoxidase precursor 245bp later and then a 168aa putative Mg^{2+} transporter MgtC/SapB (TC# 9.B.20.-.-) 23bp after that. None of these genes occur near the FUPA30 gene at high frequency.

Rpa2 (*Rhodopseudomonas palustris* CGA009 852aa, gi:39935958) is encoded as the fifth gene of a 5-gene operon which starts and proceeds downstream as follows:

1) peptidyl-prolyl cis-trans isomerase ppiD (EC 5.2.1.8) (Mg^{2+} suggested in Eukarya, but no data for prokaryotes [Jordens J, et al., 2005])
2) 37bp later, anthranilate phosphoribosyltransferase (EC 2.4.2.18)
3) 65bp later, indole-3-glycerol phosphate synthase (EC 4.1.1.48) (Ca^{2+}, Mg^{2+}, Mn^{2+} and Na^{+} "significantly affect activity, the optimal concentration is about 0.4-2.0 mM" [Yang Y, et al., 2006])
4) 11bp later, molybdenum cofactor biosynthesis protein MoaC
5) 38bp later, FUPA30

Convergently transcribed to the FUPA30 gene, 128bp away, a 158aa hypothetical protein with no significant homology is encoded and divergently from that, 239bp away, a molecular chaperone/small heat shock protein and a type IV pilus assembly PilZ protein are encoded.

Upstream of the operon containing the FUPA30 gene, 3 genes are transcribed divergently, 198bp away. The proteins these 3 genes encode are triosephosphate

isomerase (EC 5.3.1.1), 176bp later preprotein translocase subunit SecG (TC# 3.A.5.1.1) and lastly, 167bp later, CTP synthase (EC 6.3.4.2). Downstream of these are 6 hypothetical proteins with no significant homology, the first two of which are convergently transcribed. Only the 5 genes in the FUPA30-encoding operon and the 3 sharing a possible regulatory region with it are considered likely to be coexpressed.

Bps1 (*Burkholderia pseudomallei* K96243 837aa, gi:53722128) is encoded by a gene with one co-directional gene nearby on either end of it. The upstream gene encodes a 159aa putative exported protein 98bp away and the downstream gene encodes a 4 TMS, 207aa hypothetical protein 52bp away. There is a 1,026bp space between the upstream gene and the next closest gene in the upstream direction, which is divergently transcribed alone. This gene encodes a methyltransferase family protein of uncertain function. There is also a 1,429bp space between the gene immediately downstream of the FUPA30 gene and the next closest gene in the downstream direction, which is transcribed codirectionally and is followed by one additional gene in that direction. The first encodes an 86aa hypothetical protein referred to as FIG00453281 and the next, another 570bp away, encodes poly-gamma-glutamate synthesis protein/capsule biosynthesis protein CapA. There are two largely homologous proteins encoded convergently to this which each possess DnaK domains placing them in the heat-shock protein Hsp70 molecular chaperone family. The more distant one is 615aa long, the closer, encoded 4bp away, is 935aa long and encoded 295bp away from the capA gene. These genes are indicated by SEED to occur near the FUPA30 gene at much higher

frequency than any of the other aforementioned genes. One other protein is encoded upstream of these, and while it does not appear to occur near the FUPA30 gene often, it occurs near the Hsp70 genes at very high frequency and thus may be worth noting. It is transcribed codirectionally to them, overlapping the first one by 3bp, encoding a 1TMS, 191aa hypothetical protein in the DUF2780 family.

Bba4 (*Bdellovibrio bacteriovorus* HD100 825aa, gi:42522486) is encoded as the second gene in a 4-gene operon. The upstream gene encodes ribonucleotide ABC transporter ATP-binding protein 11bp away and the downstream genes encode first a small heat shock protein Hsp20 family member 57bp away followed 16bp later by rhodanese-like domain containing phage shock protein. This operon is met convergently, 14bp later by one transcribed convergently encoding 2 proteins: a CBS domain containing protein and a158aa hypothetical protein with no significant homology.

This operon probably shares a regulatory site with the divergently transcribed operon found 44bp away which encodes a putative DNA polymerase epsilon subunit (EC 2.7.7.7) (the holoenzyme requires Mg^{2+} and also Zn^{2+} [Kornberg T, et al., 1974 & Setlow P; 1974]) and 160bp later a 1 TMS, 848aa hypothetical protein Bd0913 predicted by Glimmer/Critica. This operon is met convergently, 16bp later, by one transcribed convergently encoding 4 proteins. In order of transcription, these proteins are:

1) OsmC-like protein
2) 13bp later, beta-ketoadipate enol-lactone hydrolase (EC 3.1.1.24) (no known metal cation requirements)

3) 1bp later, nucleoside-diphosphate-sugar epimerases
4) 70bp later, DnaJ domain protein/probable heat shock protein (DnaJ domain's function is docking with Hsp70).

While the operons convergently flanking the predicted transcription unit aren't found near FUPA30 with particularly high frequency, the continued presence of heat shock protein-related genes is interesting.

RegPredict analysis of the Proteobacterial FUPA30 genes was performed using the intergenic sequence preceding the first genes of each of the FUPA30-encoding operons. This ended up being the sequence preceding the FUPA30 gene itself in *B.japonicum*, while in *R.palustris* the sequence preceding the gene encoding peptidyl-prolyl cis-trans isomerase was used, in *B.bacteriovorus* the sequence preceding the ABC transporter gene was used, and in *B.pseudomallei* the sequence upstream of the co-directional, upstream gene encoding 159aa putative exported protein was also used to ensure both options were considered because of the considerable space between the genes. Both regions in *B.pseudomallei* generated operon predictions in RegPredict and may account for a greater number of coregulatory predictions in this organism than the others, which is suspected to be exacerbated by the its genome being the second largest (7.2 Mb), *B.japonicum* being the largest (9.1 Mb) and containing the second-most predictions. Conversely, *B.bacteriovorus*, which has the smallest genome by far (3.8 Mb) contained almost no regulatory predictions whatsoever besides those preceding FUPA30. Without regulatory sequence predictions for any genes in this genome besides FUPA30, these results are essentially only useful as reinforcement for predictions in other

genomes. This highly unusual result may be may be attributed to the combination of genome size, available intergenic sequence size (available sequence extracted for *B.bacteriovorus* was only 44bp long, whereas ~285bp were available for the others) and *B.bacteriovorus* having the greatest phylogenetic separation in the group.

The first predicted regulatory sequence motif to be described preceding the FUPA30 encoding operons in this group in both *B.pseudomallei* and *B.japonicum*. The sequences, preceded genes and respective distances upstream are described below:

B.pseudomallei	FUPA30 gene	41bp	CcGCCGcgC|GgaCGGCcG, score=5.92
B.japonicum	FUPA30 gene	246bp	CcGCcGgCA|TGaCaGCcG, score=6.20
		46bp	CGGCcgtCc|tGctcGCCG, score=4.82

Several related predicted regulatory sequences are found elsewhere in these genomes. The first related sequences (ccGCcGgCG|CGgCcGCcc, score=5.09; CGGCCGcCg|tGcCGGCCG, score=5.79) in this set are found in *B.pseudomallei*, 134 and 73bp upstream (respectively) of a bicistron containing *flhA* which encodes S-(hydroxymethyl)glutathione dehydrogenase (EC 1.1.1.284) (may require Zn^{2+} [Staab CA, et al., 2009]) and *fghA* which encodes S-formylglutathione hydrolase (EC 3.1.2.12) (no known metal requirements).

The second related sequence (CcGCcGcCg|aGcCcGCcG, score=5.55) in this set is also found in *B.pseudomallei*, 144bp upstream of the gene *alaS*, which encodes Alanyl-tRNA synthetase (EC 6.1.1.7) (Mg^{2+} required, "potential role of the coordinated Zn^{2+} in editing substrate specificity" [Pasman Z, et al., 2011]).

The third related sequences in this set is found in *B.japonicum* (ccGCCGcCc|tGaCGGCca, score=5.47) 105bp upstream and in *R.palustris*

(cGcCcGGCC|GGCCcGcCt, score=5.19; cGGCCGcCC|GGcCGGCCc score=5.19), 43 and 47bp of a tricistron encoding:

1. *ribBA* encodes 3,4-dihydroxy-2-butanone 4-phosphate synthase (EC 4.1.99.12) (Mg^{2+} required, Hg^{2+} and Ni^{2+} may substitute with moderate activity [Kumar P, et al., 2010]) / GTP cyclohydrolase II (EC 3.5.4.25) (requires either Mg^{2+} or Zn^{2+}, both found natively [Kaiser J, et al., 2002; Blau N, et al., 1985])
2. *gcdH* encodes Glutaryl-CoA dehydrogenase (EC 1.3.99.7) (no known metal requirements)
3. Metallo-beta-lactamase family protein 65% similar to Hydroxyacylglutathione hydrolase (glxII) (EC 3.1.2.6) (binuclear metalloenzyme with Zn^{2+}, behaves similarly with $Fe^{2/3+}$; Co^{2+} and Mn^{2+} also work but thiolate bonds are less covalent and weaker [Campos-Bermudez VA, et al., 2010; Sukdeo N, et al., 2008])

The fourth related sequence (CGGCCGGCa|aGCCGGCCG, score = 5.55) is found in *B.pseudomallei,* 107bp upstream of a 7-gene operon encoding the following proteins:

1. *hemF* encodes Coproporphyrinogen III oxidase, aerobic (EC 1.3.3.3) (Mn^{2+} required, Fe^{2+} and Cu^{2+} also found with this enzyme [Breckau D, et al., 2003; Macieira S, et al., 2003])
2. Nicotinate-nucleotide adenylyltransferase, bacterial NadD family (EC 2.7.7.18) (Mg^{2+} required [Imsande et al., 1961])
3. Iojap protein

4. LSU m3Psi1915 methyltransferase RlmH, ybeA
5. Septum formation protein Maf
6. *cafA* encodes Cytoplasmic axial filament protein CafA and Ribonuclease G (EC 3.1.4.-)
7. putative phage integrase

The fifth related sequence (CtcCCGGCG|CGCCGGccG, score=5.41) in this set is also found in *B.pseudomallei*, 77bp upstream of a tricistron encoding Type IV pilus biogenesis proteins PilMNP

The second predicted regulatory sequence preceding FUPA30-encoding operons in this group is found in 2 species as described below including upstream distance:

B.pseudomallei	*moaC*	93bp	CGCGCcgGCcG|CcGCtcGCGCG, score=5.52
	FUPA30 gene	37bp	CGcGCGgaCGG|CCGgaCGCaCG, score=5.63
R.palustris	*ppiD*	126bp	CGCgACatCgG|CtGcgGTtGCG, score=5.87

The motif set is especially interesting because *moaC* is one of the 3 genes between *ppiD* and the FUPA30 gene in *R.palustris*. Several other sequences are found which are predicted to coregulate various other operons/genes. All of them are found in *B.pseudomallei* alone. The first member of this motif set is a sequence (CGCGgcCGccG|CaaCGcgCGCG, score=5.28) found 219bp upstream of a monocistron encoding secreted microbial collagenase (EC 3.4.24.3) (Zn^{2+} found at 1:1 ratio associated with enzyme, Ca^{2+}, Co^{2+}, Mg^{2+}, Mn^{2+}, Sr^{2+} all activate, though only Ca^{2+} is found associated with native enzyme and Co^{2+} inhibits at high concentration [Bond MD, et al., 1984]).

The second related member of this motif set is a sequence (CGCGCgaGCgG|CgGCcgGCGCG, score=5.52) found 75bp upstream of a monocistron encoding zinc metalloprotease (EC 3.4.24.-).

The third related member is a sequence (CGCGACcacgG|CgagcGTCGCG, score=5.28) found 40bp upstream of a monocistron encoding cell division transporter, ATP-binding protein FtsE (TC 3.A.5.1.1).

The fourth member of this motif set is a sequence (cGCGgGggcgG|CgaggCgCGCa, score=5.32) found 67bp upstream of a tricistron encoding the following proteins:

1. heat-inducible transcription repressor HrcA
2. ferrochelatase, protoheme ferro-lyase (EC 4.99.1.1)
3. *hslR* encodes Ribosome-associated heat shock protein implicated in the recycling of the 50S subunit (S4 paralog)

The fifth member of this motif set is a sequence (CGCGCGGCCgG|CgGGCCGCGCG, score=5.33) found 92bp upstream of a monocistron encoding Manganese transport protein MntH.

The sixth and last member of this motif set is actually a pair of sequences (CGCGgcttcgG|CtcggcgCGCG, score=5.28; CGCGCgcgcgG|CgcgcgGCGCG, score=5.01) found 83 and 67bp, respectively, upstream of an operon encoding the following proteins:

1. Type III secretion inner membrane protein (T3SIM)

   (YscU,SpaS,EscU,HrcU,SsaU; homologous to flagellar export components)

2. T3SIM channel protein (LcrD,HrcV,EscV,SsaV)
3. HpaP protein; Type III secretion protein (YscP)
4. T3SIM protein (YscQ,homologous to flagellar export components)
5. T3SIM protein (YscR,SpaR,HrcR,EscR; homologous to flagellar export components)
6. T3SIM protein (YscS,homologous to flagellar export components)
7. FIG011069: putative type III secretion protein
8. T3SIM protein (YscD,homologous to flagellar export components)
9. HrpD6 putative Type III secretion protein
10. 62aa hypothetical protein with no significant homology
11. Type III secretion HpaB protein

The third predicted regulatory group is found in 3 genomes as described, including distance upstream, below:

B.pseudomallei	FUPA30 gene	37bp	CGcgCGGaC|GgCCGgaCG, score=5.32
B.japonicum	FUPA30 gene	42bp	CgTCCtGCt|cGCcGGAaG, score=5.10
R.palustris	*ppiD*	179bp	GCTGCGGcg|atCCGCAGC, score=4.75

Several other sequences are found which are predicted to coregulate various other operons/genes. All of them are found in *B.pseudomallei* in this group as well. The first member of this motif set is a sequence (CGCGCGGCG|CGCCGCGCG, score=5.59) found 141bp upstream of a monocistron encoding a Na^+:H^+ antiporter (VIMSS id=740256, locus_tag=BPSL0357).

The second member of this motif set is a sequence

(CGCGCgGCG|CGCgGCGCG, score=5.37) found 63bp upstream of the same Type III secretion system operon found in the previous group. The extremely high GC content of the region makes palindromic regulatory sequence resolution difficult to interpret.

The third member of this motif set is a pair of sequences (cgcgCgGcG|CcCgGgagc, score=5.10; CGCcCGGcC|GcCCGcGCG, score=4.97) found 22 and 146bp, respectively, upstream of the operon *gspKLMN*, which encodes the respective General secretion pathway proteins KLMN.

The fourth member of this motic set is a sequence (CGtGCGGCG|CGCCGCgCG, score=5.59) found 208bp upstream of a bicistron encoding a 125aa hypothetical protein member of DUF3175 superfamily with 51% similarity to the C-terminal half of CMP/dCMP deaminase zinc-binding protein, and another Na^+:H^+ antiporter (VIMSS id=745685, locus_tag=BPSS2210).

The fourth predicted regulatory group is also found in 3 genomes as described, including distance upstream, below:

B.pseudomallei	159aa protein gene	197bp	GCGCGGCGg|C|aCGCCGCGC, score=5.58
B.japonicum	FUPA30 gene	196bp	GcTCcGgAT|C|ATgCcGAcC, score=4.75
R.palustris	*ppiD*	107bp	GcGaCGCCa|C|aGGCGgCaC, score=4.81

Predicted regulatory sequences are found preceding other genes in each of these genomes. There are three such ssequences found in *R.palustris*. The first to be described is a sequence (GaGCGGCaT|C|AgGCCGCgC, score=5.17) found 87bp upstream of a single gene encoding carbonic anhydrase (EC 4.2.1.1) (zinc metalloenzyme, Co^{2+} can substitute for Zn^{2+} [Elleby B, et al., 2001]). The second member of this motif set found

in *R.palustris* is a sequence (GcGCGGCcT|G|AtGCCGCtC, score=5.10) found 113bp upstream of the gene *citE*, which encodes citrate lyase beta chain (acyl lyase subunit) (citE) (EC 4.1.3.6) (strictly dependent on the presence of Mg^{2+} or Mn^{2+} [Sivaraman, et al., 1979; Nilekani S, et al., 1983]). The third member found in this genome is a sequence (GctCcGCgc|C|atGCcGccC, score=5.15) found 89bp upstream of an operon encoding the following proteins:

1. Inactive homolog of metal-dependent glycoprotease (M22) metalloprotease, putative molecular chaperone
2. *rimI* encodes Ribosomal-protein-S18p-alanine acetyltransferase (EC 2.3.1.-)
3. putative manganese uptake regulation protein MUR
4. phosphatase yieH (EC 3.1.3.-)
5. miaB encodes tRNA-i(6)A37 methylthiotransferase
6. phosphate starvation-inducible ATPase PhoH with RNA binding motif
7. putative metalloprotease fusion protein
8. magnesium and cobalt efflux protein CorC
9. lnt encodes apolipoprotein N-acyltransferase (EC 2.3.1.-) / copper homeostasis protein CutE

Only one related member of this motif set is found in *B.japonicum*, a sequence (GCGCGGCgc|G|atGCCGCGC, score=5.37) 76bp upstream of a tricistron encoding a 68aa hypothetical protein with no significant homology followed by isoquinoline 1-oxidoreductase alpha and beta (small and large) subunits, CoxS/CutS homologs (EC 1.3.99.16) (Iron arranged in (2Fe-2S) clusters, Mo also found in native enzyme at 1:1

ratio [Lehmann M, et al., 1994]).

The remaining sequences in this motif set were all found only in *B.pseudomallei*. The first of these is a sequence (GcgCGGCGc|G|cCGCCGaaC, score=5.17) found 87bp upstream of a bicistron encoding predicted Na^+ symporter small subunit, and Na^+:solute symporter family protein (VIMSS Ids: 743807 and 74308, locus tags: BPSS 0376 and 0377).

The next member of this motif set is a sequence (GcGCCGCGc|A|aCGCGGCcC, score=5.00) is found 116bp upstream of a tricistron encoding spermidine synthase, putative TRAP-type C4-dicarboxylate transport system protein and another Na^+:solute symporter (VIMSS id=741002, locus_tag=BPSL1093).

The next member of this set is a sequence (GcGCcGCcc|G|acGCcGCcC, score=5.28) found 156bp upstream of a bicistron encoding two UspA proteins; one 164aa long and the other 279aa long.

The next and last member of this set is a sequence (GCGacGCGT|C|ACGCcgCGC, score=5.22) found 215bp upstream of an operon encoding the following proteins:

1. High-affinity branched-chain amino acid transport system permease protein LivH (TC 3.A.1.4.1)
2. Branched-chain amino acid ABC tranpsorter, permease/ATP-binding protein
3. Branched-chain amino acid transport ATP-binding protein LivF (TC 3.A.1.4.1)
4. *gidA* encodes tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA

5. *gidB* encodes rRNA small subunit methyltransferase, glucose inhibited division protein GidB
6. Chromosome (plasmid) partitioning protein ParA / Sporulation initiation inhibitor protein Soj
7. Chromosome (plasmid) partitioning protein ParB /Stage 0 sporulation protein J
8. Arsenical pump membrane protein

The fifth predicted regulatory group is found in 2 genomes as described, including distance upstream, below:

B.pseudomallei	159aa protein gene	47bp	CGCcgcCGT|G|ACGcgcGCG, score=5.37
	FUPA30 gene	46bp	CGatCcCGC|C|GCGcGgaCG, score=4.88
R.palustris	*ppiD*	167bp	cGCAGCatC|G|GccGCTGCa, score=4.97

Only one of the corresponding predicted regulatory sequences in this motif set is found in *R.palustris*. The sequence (CGCGCCCGT|G|ACGGGCGCG, score=5.24) is found 47bp upstream of a 4-gene operon encoding the following proteins:

1. putative short-chain alcohol dehydrogenase
2. Tetratricopeptide repeat (TPR) protein
3. acetyl-coenzyme A synthetase (EC 6.2.1.1) (Mg^{2+} required for activation, Mn^{2+} can fully replace at 60% higher optimal concentration, Ca^{2+} and Co^{2+} can each replace with 50% efficiency at optimal concentration [Preston GG, et al., 1990])
4. 86aa hypothetical protein with no significant homology

The remaining sequences in this motif set were all found only in *B.pseudomallei*. The first of these is a sequence (CGCcGcCcC|G|GcGcCcGCG, score=5.25) is found 91bp upstream of a 4-gene operon encoding the following proteins:

1. *citE* encodes citrate lyase beta chain (EC 4.1.3.6) (strictly dependent on the presence of Mg^{2+} or Mn^{2+} [Sivaraman, et al., 1979; Nilekani S, et al., 1983])
2. Probable signal peptide protein
3. *prpD* encodes 2-methylcitrate dehydratase (EC 4.2.1.79) ("possesses 1 unstable iron-sulfur center per monomer, absolutely required for activity" [Grimek TL, et al., 2004])

citB encodes aconitate hydratase (EC 4.2.1.3) (binds 1 4Fe-4S cluster per subunit [Tang Y, et al., 2005] and possibly Mg^{2+} [Tsuchiya D, et al., 2008]) / 2-methylisocitrate dehydratase (EC 4.2.1.99) (no known metal requirements)

The second member of this motif set is a sequence (gGCcgCCGt|G|gCGGgcGCt, score=5.24) found 242bp upstream of a tricistron encoding the following proteins:

1. aspartate/tyrosine/aromatic aminotransferase
2. *hom* encodes homoserine dehydrogenase (EC 1.1.1.3) (no metal requirements found for prokaryotes)
3. *thrC* encodes threonine synthase (EC 4.2.3.1) (no known metal requirements)

The third member of this motif set is a sequence (CGcggcgGC|G|GCgcggaCG, score=5.32) found 222bp upstream of a monocistron encoding threonine dehydratase biosynthetic (EC 4.3.1.19) (no known metal requirements in prokaryotes).

The fourth member of this motif set is a sequence (CGCttcgtC|G|GcgcgcGCG,

score=5.32) found 165bp upstream of a monocistron encoding LysE family putative threonine efflux transporter.

The fifth member of this motif is actually a pair of sequences (CaCGccCGc|G|aCGccCGcG, score=4.83; CGCcGcCGc|C|aCGcCcGCG, score=5.37) found 154 and 163bp upstream, respectivley, of a 4-gene operon encoding the following proteins:

1. 260aa hypothetical protein member of DUF3348 superfamily
2. Methyl-accepting chemotaxis protein
3. Chemotaxis motB protein, OmpA flagellar motor protein family
4. 280aa hypothetical protein member of DUF2894 superfamily

The sixth and last member of this motif set, as well as this proteobacterial FUPA30 coregulatory analysis group, is a pair of sequences (CGCggcCGC|G|GCGcggGCG, score=5.68; CGCCGCcGC|G|GCcGCGGCG, score=5.46) found 175 and 181bp upstream, respectively, of an operon encoding the following proteins:

1. *narG* encodes Respiratory nitrate reductase alpha chain, NapA homolog (EC 1.7.99.4) (requires Fe and Mo both for efficient electron transfer because “the diheme electron transfer small subunit NapB binds to the large subunit with heme II in close proximity to the [4Fe-4S] cluster of NapA. The plasticity of the complex contributes to an efficient electron transfer in the complex from the heme I of NapB to the molybdenum catalytic site of NapA” [Arnoux p, et al., 2003])

2. *narH* encodes Respiratory nitrate reductase beta chain, NapB homolog (EC 1.7.99.4)
3. *narJ* encodes Respiratory nitrate reductase delta chain, NapD homolog (EC 1.7.99.4)
4. *narI* encodes Respiratory nitrate reductase gamma chain, NapC homolog (EC 1.7.99.4)
5. Peptidyl-prolyl cis-trans isomerase ppiD (EC 5.2.1.8) (Mg^{2+} suggested in Eukarya, but no data for prokaryotes [Jordens J, et al., 2005])
6. *narK* encodes Nitrate/nitrite transporter

TC# 3.A.3.30.2 contains FUPA30s found in Bacteroidetes and is represented by **Fjo1** (*Flavobacterium johnsoniae* UW101 838aa, gi:146302203), which is encoded codirectionally 154bp downstream of a large operon and is likely part of it. The operon encodes proteins as follows in order of transcription:

1) 1,446aa cobalt-zinc-cadmium resistance protein CzcA; Cation efflux system protein CusA (TC# 2.A.6.1.-)
2) 7bp later, probable Co/Zn/Cd efflux system membrane fusion protein (TC# 2.A.6.1.-)
3) -10bp later (overlap) diacylglycerol kinase (EC 2.7.1.107) ("enzyme requires a free divalent metal cation: Mg^{2+}, Mn^{2+}, Co^{2+}, Cd^{2+} or Zn^{2+}" [Walsh JP, et al., 1992])
4) 14bp later, sulfatase family protein

5) 71bp later, thioredoxin disulfide isomerase
6) -3bp later (overlap), thiamin biosynthesis lipoprotein ApbE
7) -3bp later, putative arginine decardoxylase
8) 33bp later, 392aa hypothetical protein with no significant homology
9) 2bp later, PAP2 superfamily protein
10) 4bp later, NiFe-hydrogenase I cytochrome b subunit
11) 33bp later, nodulin 21-related protein
12) 79bp later, metal-dependent phosphohydrolase, HD subdomain
13) 89bp later, tetracycline resistance element mobilization regulatory protein rteC
14) 154bp later, FUPA30

All of these upstream genes appear to occur with medium to high frequency near the FUPA30 gene except that encoding the sulfatase family protein. While one more gene is transcribed codirectionally to this operon upstream encoding amuramoyltetrapeptide carboxypeptidase (EC 3.4.17.13), it is 2.1kb away and thus presumed not to be cotranscribed.

The next gene downstream of the FUPA30 gene is convergently transcribed 236bp away and encodes a 761aa protein which is highly homologous to the C-terminal end of ~1.8kD N6 DNA methylase. It is found 747bp away from the next 2 genes upstream of it, which are codirectionally transcribed and encode a 92aa hypothetical protein referred to as FIG00654054 followed 2bp later by one encoding a probable P-loop-containing nucleoside triphosphate hydrolase. The remaining genes found in this direction are either oriented or distanced such as to make coexpression ever more

unlikely, and most encode hypothetical proteins with no significant homology. However, two genes are worth noting despite this, both due to their high frequency of occurrence near the FUPA30 gene. The nearer one encodes a 502aa protein highly homologous to the C-terminal half of the PAS/PAC sensor signal transduction histidine kinases with TC# 2.A.21.9.-, which include nitrogen and proline sensor-receptor domains. The farthest gene of interest in the gene neighborhood encodes a homologue of the CzcA/CusA protein at the beginning of the FUPA30-encoding operon, but this protein appears to be only 1037aa in length. The 410 residue difference in length of these two proteins is accounted for by a TolC domain at the C-terminus of the longer one.

RegPredict analysis of FUPA30 co-regulation in Bacteroidetes was performed using intergenic sequence preceding genes from the *Flavobacterium johnsoniae* putative operon described above encoding FUPA30, 392aa hypothetical protein, and CusA/CzcA. Intergenic sequence preceding the closest orthologues, found in *Chryseobacterium gleum* ATCC 35910, were also used to expand the training set. In addition to *F.johnsoniae*, the genomes of *F.psychrophilum* JIP02/86, *Candidatus Sulcia muelleri* SMDSEM and *Flavobacteriaceae bacterium* 3519-10 were also examined for comparison, although none of them possesses a FUPA30 gene.

The first predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *F.johnsoniae* at two loci as described below including upstream distance:

156bp	392aa hypothetical protein gene	CgATgCCga|A|atGGtATtG, score=4.41
79bp	FUPA30 gene	TTTCtCtgc|G|ttcGtGAAA, score=4.01

Four similar sequences are predicted to be members of this motif set. The first sequence (ggAtgCggt|A|ttgGttTaa, score=4.33) is found 198bp upstream of a tricistron encoding H.8 outer membrane protein precursor, quinol-dependent nitric-oxide reductase (EC 1.7.99.7) (no known metal requirements), and nitric oxide-dependent regulator DnrN/NorA.

The second member in this motif set is a sequence (TgTtTCtgt|A|ttgGAgAaA, score=4.32) found 209bp upstream of an operon encoding the following proteins:

1. *hemE* encodes Uroporphyrinogen III decarboxylase (EC 4.1.1.37) (possible slight activation with Mn^{2+} [Jones RM, et al., 1993])
2. 5 TMS, 225aa hypothetical protein with no significant homology
3. 125aa hypothetical protein with no significant homology
4. Ribosomal-protein-L7p-serine acetyltransferase
5. *hemF* encodes Coproporphyrinogen III oxidase, aerobic (EC 1.3.3.3) (Mn^{2+} required, Fe^{2+} and Cu^{2+} also found with this enzyme [Breckau D, et al., 2003; Macieira S, et al., 2003])

The third member of this motif set is a sequence (gAATgcTGA|A|ATGaaATTt, score=3.77) found 152bp upstream of an operon encoding the following proteins:

1. Deoxyguanosinetriphosphate triphosphohydrolase (EC 3.1.5.1) (Mg^{2+} required for activity, Mn^{2+} and Co^{2+} can substitute with 30 and 17% activation rate, respectively [Seto D, et al., 1988])

2. *lpxD* encodes UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.-)
3. *lpxC/fabZ* encodes N-acetylglucosamine deacetylase (EC 3.5.1.-) / (3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase (EC 4.2.1.-)
4. *lpxA* encodes Acyl-[acyl-carrier-protein]--UDP-N-acetylglucosamine O-acyltransferase (EC 2.3.1.129) (Mg^{2+} is required for stabilization of substrate acyl-carrier protein in a helical conformation, Ca^{2+}, Mn^{2+} [Gong H, et al., 2007])
5. *efp* encodes translation elongation factor P
6. UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.-)
7. 120aa hypothetical protein with no significant homology
8. *sucD* encodes Succinyl-CoA ligase [ADP-forming] alpha chain (EC 6.2.1.5) (Mg^{2+} required, the true substrate is the MgADP-complex [Joyce MA, et al., 2000])
9. 142aa hypothetical protein with no significant homology

This operon is also found to be preceded by a related sequence in *F.psychrophilum*, but is missing the last hypothetical protein gene.

The fourth member of this motif set is a sequence (cgATtTcgT|A|AttAtATtt, score=4.34) found 207bp upstream of a monocistron encoding transcriptional regulator of rhamnose utilization, GntR family.

The next sequence predicted to regulate the FUPA30 gene is found preceding the CzCA/CusA gene and the FUPA30 gene itself. The sequences

(ATAaATACtG|CtGTATcTAT, score=5.09; AGCTgTTaTc|aAgAAtAGCT, score=4.43) are found 62 and 61bp upstream of the respective genes.

The first member in this motif set is a sequence (tTaAtTcatg|atgtAtTgAt, score=4.66) found 202bp upstream of a bicistron which encodes a cell division inhibitor and deoxyribodipyrimidine photolyase (EC 4.1.99.3) (does not require divalent cations [Sancar A, et al., 1984]).

The second member in this motif set is a sequence (aatatTcCtg|atGaAtataa, score=4.47) found 200bp upstream of a tricistron encoding plectin 1 isoform 8, putative TonB-dependent siderophore receptor, TonA, a putative adhesin/internalin precursor SprB/gliding motility-related protein with FlgD domain and 1TMS, 333aa Bacteroidetes-specific putative membrane protein 39.7% similar to C-terminus of TonB-dependent receptor. In *F.psychrophilum*, a related sequence precedes a related operon which also includes the gene *mltD* that encodes peptidoglycan-binding LysM:SLT.

The third member in this motif set is a sequence (aTtAtTtaAg|aTgcAtTtAa, score=5.01) found 67bp upstream of a 4-gene operon, which encodes the following proteins:

1. TonB-dependent receptor, plug protein
2. RagB/SusD domain protein; putative outer membrane protein probably involved in nutrient binding
3. COG3866 Pectate lyase
4. 221aa hypothetical protein in DUF3826 superfamily, putative sugar-binding proteins

The fourth member in this motif set is a sequence (TTttcTTCtg|atGAAttgAA, score=4.57) found 196bp upstream of a 4-gene operon which encodes the following proteins:

1. putative plasmid stabilization system protein
2. putative Holliday junction resolvase
3. malate:quinone oxidoreductase (EC 1.1.99.16) (no known metal requirements)
4. isochorismatase family protein

The fifth member in this motif set is a sequence (TTCAgTAaAg|aTgTAtTGAA, score=4.65) found 100bp upstream of a 4-gene operon encoding the following proteins:

1. FMN reductase, NADPH-dependent
2. *sufC* encodes Iron-sulfur cluster assembly ATPase protein SufC
3. S23 ribosomal
4. *sufD* encodes Iron-sulfur cluster assembly protein SufD

The third predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *F.johnsoniae* at two loci as described below including upstream distance:

204bp	392aa hyp. protein gene	gctTGGTtCc|tGcACCAcag, score=4.05
75bp	FUPA30 gene	tCtGCgTTCg|tGAAaGCtGt, score=4.32

The operons preceded by this motif set are exceptionally rich in transcriptional regulators.

The first member of this motif set is a sequence (TaAtAATTcg|ttAATTtTaA, score=4.04) found 210bp upstream of a 6-gene operon encoding the following proteins:

1. Peptidase M23B
2. transcriptional regulator, MerR family
3. LemA family protein
4. FIG004694: Hypothetical protein of unknown function DUF477
5. Beta-propeller domains of methanol dehydrogenase type
6. Predicted transcriptional regulator of sulfate adenylyltransferase and sulfate transport, Rrf2/BadM family

The second member of this motif set is a sequence (tgAgcgTaCa|gGgAtttTtt, score=4.17) found 55bp upstream of another 6-gene operon, this one encoding:

1. 269aa hypothetical protein with no significant homology
2. RTX toxins and related Ca^{2+}-binding proteins
3. 303aa bacteroidetes-specific putative membrane proteinOmpA/MotB domain protein
4. 4TMS, 327aa hypothetical protein with no significant homology
5. 230aa hypothetical protein 40% similar to envelope protein EnvF from *Salmonella enterica subsp. enterica* serovar Heidelberg str. 41573, length=235

The third member of this motif set is a sequence (tcAtcATaaA|TgaATctTtt, score=4.08) found 55bp upstream of a tricistron encoding two copies of transcriptional regulator, AraC family and one putative TonB-dependent receptor.

The fourth member of this motif set is a sequence (tcAggaataC|GgcacttTat, score=4.01) found 156bp upstream of a bicistron encoding transcriptional regulator, AraC family and tetracycline resistance element mobilization regulatory protein rteC.

The fifth member of this motif set is a pair of sequences (tcttcAatAg|tTgaTttttt, score=3.95; gcagatTTCG|CGAAtctttt, score=4.37) found 227 and 180bp upstream, respectively, of a monocistron encoding LysR family regulatory protein CidR.

The sixth and last member of this motif set is a sequence (gcagAaTtCC|GGgAaTtgaa, score=4.03) found 204bp upstream of a monocistron encoding Molybdenum cofactor biosynthesis protein MoaA.

The fourth predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *F.johnsoniae* at two loci as described below including upstream distance:

63bp	CzcA/CusA gene	gATAaATACtG|CtGTATcTATt, score=4.69
62bp	FUPA30 gene	aAGCTgTTaTc|aAgAAtAGCTc, score=4.58

Four similar sequences are predicted to coregulate various genes/operons within this genome. The first of these is a sequence (aTTAaATTata|ataAATaTAAc, score=4.57) found 71bp upstream of a 4-gene operon encoding the following proteins:

1. 6-phosphofructokinase (EC 2.7.1.11) (requires Mg^{2+}, Mn^{2+}, Co^{2+}, Zn^{2+} and Ni^{2+} may substitute with respective activities of 75, 50, 35 and 12% [ForDyce AM, et al., 1982])

2. *gapA3* encodes NAD-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) (no known divalent cation requirements)
3. N-acetylglucosamine kinase-like protein
4. *mgsA* encodes Methylglyoxal synthase (EC 4.2.3.3) (no known cation requirements)

The second member of this motif set is a sequence (aaTtAtTtaAg|aTgcAtTtAaa, score=4.58) found 68bp upstream of a 4-gene operon previously described in the second motif set, encoding the following proteins:

1. TonB-dependent receptor, plug protein
2. RagB/SusD domain protein; putative outer membrane protein probably involved in nutrient binding
3. COG3866 Pectate lyase
4. 221aa hypothetical protein in DUF3826 superfamily, putative sugar-binding proteins

The third member of this motif set is a sequence (AaTAttTTaTg|aAaAAttTAaT, score=4.62) found 89bp upstream of a bicistron encoding protease II (EC 3.4.21.83) (Ca^{2+} stimulates amidase and proteolytic activities, Co^{2+}, Fe^{2+}, Hg^{2+} and Zn^{2+} are inhibitory [Pacaud M, et al., 1975]) and cystathionine beta-synthase (EC 4.2.1.22) (may contain Ca^{2+}, Fe^{2+} or Zn^{2+} [Proudfoot M, et al., 2008]).

The fourth and last member of this motif set is a sequence (AAaATaTTatg|ataAAaATaTT, score=4.59) found 86bp upstream of a monocistron encoding transcriptional regulator, TetR family.

The last predicted regulatory sequence for this group has only one sequence (ctCtGCgTTCg|tGAAaGCtGtt, score=4.62) preceding the FUPA30 gene by 76bp. Additionally, only one member is predicted for this set. The sequence (gtAAtagTTtc|agAAtctTTtt, score=4.94) is found 222bp upstream of an operon encoding the following proteins:

1. *ahcY* encodes Adenosylhomocysteinase (EC 3.3.1.1) (no known metal requirements)
2. 94aa hypothetical protein with no significant homology
3. 1 TMS, 359aa hypothetical protein
4. Ethidium bromide-methyl viologen resistance protein EmrE
5. ADA regulatory protein / Methylated-DNA--protein-cysteine methyltransferase (EC 2.1.1.63) (may require Zn^{2+} [Guengerich FP, et al., 2003])
6. prolyl 4-hydroxylase, alpha subunit
7. 81aa hypothetical protein containing Zn/metal binding domain of Ada
8. DNA repair system specific for alkylated DNA

TC# 3.A.3.30.3 contains FUPA30s found in the Spirochaetes and is represented by **Lbi8** (*Leptospira biflexa serovar* Patoc strain 'Patoc 1 (Paris)' 837aa, gi:189911895), which is encoded at the beginning of a 4-gene operon and divergent to one other gene. The three genes downstream of the FUPA30 gene encode a putative patatin-like phospholipase 24bp away, followed by two copies of PAS/PAC sensor signal

transduction histidine kinase similar to the one described above in *F.johnsoniae* encoded with 39 and 44bp intergenic spaces preceding the genes. The gene divergent to the FUPA30 gene is 110bp away and encodes a putative C69 superfamily U34 dipeptidase which is 58% homologous to peptidase C45 acyl-coenzyme A:6-aminopenicillanic acid acyl-transferase. The described transcription unit is flanked on both sides by convergently transcribed genes. At the end of the operon containing the FUPA30 gene a single gene encoding a serine phosphatase RsbU/regulator of sigma subunit is found 4bp away from the terminal gene of the operon. This serine phosphatase gene is in turn transcribed divergently from a bicistron which begins 158bp away and encodes ClpB protein and, 72bp later, 2,5-diketo-D-gluconic acid reductase A, an oxidoreductase of aldo/keto reductase family, subgroup 1. The U34 dipeptidase gene has a 4-gene operon transcribed convergently to it that ends with its last gene overlapping that of the U34 dipeptidase gene by 23bp. This operon encodes first another serine phosphatase RsbU/regulator of sigma subunit followed, 294bp later, by a two-component sensor histidine kinase, a pyridine nucleotide-disulphide oxidoreductase domain protein 42bp after that and finally D-glycerate 2-kinase (EC 2.7.1.-) 5bp after that. This operon also has a single gene transcribed divergently from it, 39bp away, but it encodes a 47aa hypothetical protein with no significant homology to known proteins and is found with only very low frequency near the FUPA30 gene. The genes encoding ClpB and the two-component sensor histidine kinase are also found nearby at only low frequency and thus should receive little weight in determining FUPA30 protein function.

RegPredict analysis is only available for 2 species of Spirochaetes, both in the Treponema genus. These lack a FUPA30 gene and thus will not be analyzed.

TC# 3.A.3.30.4 contains FUPA30s found in Cyanobacteria and is represented by **Ava15** (*A.variabilis* ATCC 29413 867aa, gi:75910290), which is encoded as the third gene in a 3-gene operon with one gene divergently transcribed from the beginning of the operon, presumably sharing a regulatory sequence, 249bp away. The two upstream genes in the FUPA30-encoding operon encode a two-component transcriptional regulator, winged helix family protein, then, 119bp later, a two-component sensor histidine kinase is encoded, ending 1bp before the FUPA30 gene begins. 340bp further upstream is Ava14 (Family 1 (Na^+,K^+) P-type ATPase; 971aa, gi:75910286), also codirectional. The FUPA30 gene is met convergently 234bp later by the last gene in a 6-gene operon, which encodes (in order of transcription) two putative phosphoribosyl- transferases, 25bp apart, a 54aa hypothetical protein with no significant homology 191bp later, sigma 54 modulation protein/ribosomal protein S30EA 64bp after that, then a 136aa putative enzyme of poly-gamma-glutamate biosynthesis (capsule formation), CapA-like protein which is highly homologous to the C-terminus of the ~325aa complete protein (like that encoded near the FUPA30 in *B.pseudomallei*) and lastly, 18bp later, a putative methyl-accepting chemotaxis protein.

The gene divergently transcribed from the FUPA30-encoding operon encodes PepSY-associated TM helix protein. The next gene downstream of it is 485bp away and convergently transcribed. The intergenic distance and orientation of the gene would

typically make it appear of little interest, but because it encodes Ava14, it is mentioned here. Interestingly, on either end of the full set of genes just described in this organism, there are pentapeptide repeat family proteins encoded, 466bp upstream of the Family 1 ATPase gene and 187bp upstream of the phosphoribosyltransferase pair. Both of these genes are found at low frequency near the FUPA30 gene and the others, but their presence and orientation may give clues to how this gene cluster evolved in *A.variabilis* and are thus worth noting.

RegPredict analysis of the FUPA30 co-regulation in Cyanobacteria was performed using intergenic sequence upstream of the two component transcriptional regulator, winged helix family protein (txnlRWH) gene and Two-component Sensor Histidine Kinase (2ComHisK) gene found immediately upstream of the FUPA30 gene. In addition to *A.variabilis*, the genomes of *Synechococcus elongatus* PCC 7942, *Thermosynechococcus elongatus* BP-1 and *Nostoc sp*. PCC 7120 were also examined for comparison, although none of them possesses a FUPA30 gene.

The first predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *A.variabilis* at two loci as described below including upstream distance:

90bp txnlRWH gene CTtcAtTCAtg|tcTGAtTttAG, score=7.22 116bp 2ComHisK gene aTAAaAgCAgg|gtTGtTgTTAg, score=6.01

Three related sequences are predicted to coregulate other genes/operons within this genome, as well as being found identically in *Nostoc sp*. PCC 7120. The first such

sequence (CTtcaATcAgg|tgTcATctcAG, score=5.04) is found 108bp upstream of a 5-gene operon encoding the following proteins:

1. NAD(P)H-quinone oxidoreductase chain 1
2. NAD(P)H-quinone oxidoreductase chain I (EC 1.6.5.3) (contains 9 Fe-S clusters and 5-10 mM $CaCl_2$ is optimal for activity [Flemming D, et al., 2005; Sazanov LA, et al., 2003])
3. NAD(P)H-quinone oxidoreductase chain J (EC 1.6.5.3)
4. NADH dehydrogenase subunit 4L
5. NAD kinase (EC 2.7.1.23) (several divalent cations satisfy the metal ion requirement: most effective are Mn^{2+}, Mg^{2+} and Ca^{2+}; Fe^{2+}, Zn^{2+} and Co^{2+} also work- . In eukaryotes, Ca^{2+}/Calmodulin are essential [Zerez CR, et al., 1986; Kawai S, et al., 2008])

The second member of this motif set is a sequence (cTtAAttcAag|tcTagtTTtAt, score=5.16) found 29bp upstream of a bicistron, which encodes serine-protein kinase rsbW (EC 2.7.11.1) (Mg^{2+} or Mn^{2+} required for efficient activity [Sharma K, et al., 2004], no activity with Mg^{2+} alone, Ca^{2+} can partly substitute Mn^{2+}, no activity with Cu^{2+} or Zn^{2+} [Gopalaswamy R, et al., 2004]) and signal transduction histidine kinase.

The third member of this motif set is a sequence (CTtAaagCAAc|tTTGaagTtAG, score=5.13) found 136bp upstream of a monocistron encoding two component transcriptional regulator, LuxR family.

The second predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *A.variabilis* at two loci nearly identical to the previous two, but with a different line of symmetry in the palindromes and different predicted coregulated genes, still worth noting. The sequences are described below including upstream distance:

92bp txnlRWH gene tActTCAttCA|TGtcTGAttTt, score=7.12

118bp 2ComHisK gene AAatAaAAgCa|gGgTTgTtgTT, score=5.79

Four related sequences are predicted to coregulate other genes/operons within this genome, the first 3 of which are also found identically in *Nostoc sp*. PCC 7120. The first such sequence (AACtaaAtttA|TgtcTgatGAA, score=5.07) is found 211bp upstream of a bicistron encoding methyl-accepting chemotaxis protein and chemotaxis protein cheA (EC 2.7.3.-).

The second member of this motif set is a sequence (AAttTaAAttC|TgtTTgAttTT, score=5.14) found 171bp upstream of a tricistron encoding Cyanobacterial SigF-related sigma-37 type sigma factor (sporulation transcription factor in Bacillus, encoded by putative "*sigF*") and we as two hypothetical proteins with no significant homology, 281aa and 82aa in length.

The third member of this motif set is a sequence (tgcttaAttCa|aGtcTagtttt, score=5.06) found 31bp upstream of a bicistron described in the previous set which encodes serine-protein kinase rsbW (EC 2.7.11.1) and signal transduction histidine kinase.

The fourth member of this motif set is a sequence (tAAgaaAgtca|gattTgatTTt, score=5.02) found 184bp upstream of a monocistron encoding NAD-reducing hydrogenase subunit HoxF (EC 1.6.5.3) (contains 9 Fe-S clusters and 5-10 mM $CaCl_2$ is optimal for activity [Flemming D, et al., 2005; Sazanov LA, et al., 2003])

The third predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *A.variabilis* at two loci as described below including upstream distance:

81bp txnlRWH gene tGTCtgaTtT|T|AgActtGACt, score=7.01

71bp 2ComHisK gene GGTTAGaTAt|T|tTAaCTAACC, score=5.77

Four related sequences are predicted to coregulate other genes/operons within this genome, two of which are also found with high similarity in *Nostoc sp*. PCC 7120. The first such sequence (tttTtgaTtg|T|atActtAcct, score=5.03) is found 180bp upstream of a monocistron encoding Potassium channel protein TrkA. A closely related sequence is found in *Nostoc sp.* with score=4.50.

The second member of this motif set is a sequence (ttTttgaatt|T|tgaattgAtt, score=5.03) found 59bp upstream of a bicistron encoding MoxR ATPase/Mg-chelating protein and 3 TMS, 436aa hypothetical protein with von Willebrand factor A domain.

The third member of this motif set is a sequence (atTCttaatt|T|tgaattGAct, score=4.70) found 105bp upstream of a monocistron, *dnaK*, which encodes Chaperone protein DnaK/Heat shock protein Hsp70. A closely related sequence is found in *Nostoc sp.* with score=5.11.

The fourth member of this motif set is a sequence (ttttttaTtt|T|ttActtgcct, score=5.03) found 88bp upstream of a four gene operon which encodes Sulfate and thiosulfate binding protein CysP, Sulfate transport system permease proteins CysT and CysW, and Ferredoxin.

The fourth predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *A.variabilis* at two loci as described below including upstream distance:

130bp txnlRWH gene actTGGaAag|A|tgTaCCAcca, score=6.91

71bp 2ComHisK gene GGTTAGaTAt|T|tTAaCTAACC, score=5.54

Three related sequences are predicted to coregulate other genes/operons within this genome, all of which are also found with high similarity in *Nostoc sp*. PCC 7120. The first such sequence (GGTTaGaTat|A|taAaCgAACC, score=4.99) is found 45bp upstream of a tricistron encoding Nitrogen regulatory protein P-II, Glutathione S-transferase and putative nuclease protein.

The second member of this motif set is a sequence (tctTggaAAt|T|tTTagtAgcc, score=4.83) found 28bp upstream of a bicistron encoding priming glycosyltransferase involved in lipopolysaccharide synthesis and UDP-glucose 4-epimerase (EC 5.1.3.2) (contains none of the common metal ions [Arabshahi A, et al., 1988]).

The third member of this motif set is a sequence (aatTggagAt|A|tTtattAcca, score=5.08) found 57bp upstream of a monocistron encoding another serine/threonine

kinase (EC 2.7.11.1) other than that described in the first two sets in this group and four times larger than it.

The fourth and last predicted regulatory sequences preceding genes of the FUPA30-encoding putative operon are found in *A.variabilis* at two loci as described below including upstream distance:

183bp txnlRWH gene aTGtGcTac|T|taAcCtCAa, score=6.67

70bp 2ComHisK gene GTTAGaTAt|T|tTAaCTAAC, score=5.50

Two related sequences are predicted to coregulate other genes/operons within this genome, both of which are also found with high similarity in *Nostoc sp*. PCC 7120. The first such sequence is actually two sequences (cTTtGaTAt|T|tTAgCtAAa, score=5.06; aTcaaaTaA|T|TaAactaAa, score=4.63), found 45bp upstream of the first gene and 39bp upstream of the third gene, respectively, in a tricistron encoding Two-component response regulator, purine-binding chemotaxis protein, and methyl-accepting chemotaxis protein.

The second member of the motif, and last in this coregulatory analysis group, set is a sequence (atctGaTAt|T|tTAaCtaca, score=4.98) found 41bp upstream of a monocistron encoding 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase (EC 1.17.7.1) (contains $[4Fe\text{-}4S]^{2+}$ cluster [Okada K, et al., 2005]).

Several patterns of co-localized and coregulated genes come to light for the FUPA30 family, all largely in support of the characterization of this family as a

specialized calcium transporter. One of the described FUPA30 proteins, that of *B.bacteriovorus*, is referred to as PacL and is putatively characterized as a Ca^{2+} transporter based on its homology to the confirmed Ca^{2+} transporter PacL in *Synechococcus elongatus* PCC 7942 (Berkelman T, et al., 1994). Furthermore, the roles of calcium in bacteria have been largely characterized (Norris V, et al., 1996). These may be summed up into categories such as

1) an indispensible inducer of the tumbling mechanism of chemotaxis
2) a protein function modulator
3) differentiation such as sporulation, gliding and heterocyst formation
4) coordinated general reset mechanism, e.g., activation/dismissal of teams of transcriptional regulator

These functions are supported to appreciable degrees by the assortment of predicted coregulated genes described herein. Chemotaxis is supported by numerous methyl-accepting chemotaxis proteins and especially CheA, one of the proteins responsible for enacting calcium's specific role in the process. Protein function modulation is supported first and foremost by the wealth of coexpressed DnaJ, DnaK/Hsp70, Hsp20, HrcA, HslR proteins. This is especially the case for DnaK, which possesses a calmodulin-like active site domain and shows a 10-fold increase in autophosphorylation activity in the presence of calcium, as well as for NAD kinase (EC 2.7.1.23), whose eukaryotic function requires the presence of Ca-charged calmodulin. This modulation may, for example, affect progression through the cell cycle. Differentiation is best supported for gliding by the coexpression of SprB/FlgD and

OmpA/MotB. Heterocyst formation is also supported because this differentiation occurs when nitrates are removed, such as by the various respiratory nitrate reductases and various other NO-type compound reductases considered alongside the fact that the heterocyst frequency is regulated by calcium concentration. Sporulation is less well supported because while numerous sporulation-related genes were found, e.g., ParAB and SigF, they appear to all be homologues since the organisms in which they were found are not known to form spores. This observation may indicate that *some* differentiation proteins in various organisms, which are involved in radically different types of differentiation, may be phylogenetically related. Finally, the idea that calcium may serve as a coordinated general reset mechanism, for example in the activation/dismissal of teams of transcriptional regulator, originally put forth by Norris V, et al. (1996) is supported essentially by the abundance of transcription factors predicted to be coexpressed with the FUPA30 gene, but not each other. The list of transcription factors includes LysE, LysR, Crp/Fnr, GntR, MerR, AraC, TetR and LuxR. Additionally, several other genes seem to be coexpressed with FUPA30 genes, although their functions are less clear. These include cell division protein FtsE and septum formation protein Maf, as well as CzcA/CusA and CorC, $Co^{2+}/Zn^{2+}/Cd^{2+}$ and Mg^{2+}/Co^{2+} efflux proteins, respectively, manganese uptake proteins, Na^{+}:H^{+} antiporters and Na^{+}:symporters. The first two support the suggestion that FUPA30 may play a role in calcium influx relating to cell division, especially in proteobacteria, where the putative regulatory sequences are very GC-rich (see above), since calcium binds preferentially to these regions (Norris V, et al., 1996). The common role these proteins share is clearly

ion transport and consequently also membrane potential regulation as well. As such, we propose that divalent cation concentration regulation, especially where calcium is concerned - restoring resting concentration following an ion flux used to effect one of the various functions mentioned above - is the role of co-regulation of these proteins with FUPA30.

The FUPA31 ATPase Family

FUPA31 is a family of Type I P-type ATPases with 2 subfamilies, the first representing predicted functional enzymes and the second representing a particular coding sequence predicted to be a pseudogene. Those analyzed here range in size from 673-925 aas and have topologies of 2-6 TMS according to TCDB, although several high hydropathy, low amphipathicity regions in the low-TMS predicted proteins tend to occur where the seemingly missing 3rd and 4th TMS are expected to be absent from. This may reduce the range to 4-6 if those regions are in fact TMS regions as well. The first two TMS, which correspond to the standard 3 & 4 notation, cluster within a 75 aa range found between residues 30-325, 3 & 4 (which correspond to TMS 5 & 6 in the standard P-type ATPase topology) cluster within a 40 aa range about 150 residues away from 1 & 2. TMS 5 & 6 cluster together within a 42 aa range about 280 residues away from 3 & 4 and 20-60 residues from the end of the protein in TC 3.A.3.31.1 or 300 residues from the end in TC 3.A.3.31.2. The family's nearest known is Family 5 (TC 3.A.3.5.-), subfamilies 4 and 15 as well occasionally as Family 6 (TC 3.A.3.6.-), especially subfamily 17, but in some cases, FUP32 comes up first. This is expected given their phylogenetic placement approximately equidistant from these two known families but having more in common with Family 5 in regards to functional motif conservation (Chan, et al., 2010). This FUPA family is only shown in *Methylococcus capsulatus str. Bath,* 3 proteins in Chan et al., 2010. Here, *Methylocella silvestris* BL2 and *Beijerinckia indica subsp. indica* ATCC 9039 in order to expand the breadth of discussion. Mca12 is the only one of the 3 found to have FUPA31 homologues in other species, including two

which are co-localized in *Methylocella silvestris* BL2. Mca10 and Mca 11 are found to be more closely homologous to FUPA32s in other species.

RegPredict does not contain Methylococcales, Rhizobiales and as such RegPredict analysis is not possible for these genomes.

Mca10 is predicted to be a pseudogene due to its abnormal length resulting from a C-terminal fusion with an EcsC superfamily protein normally found in an operon with the ABC transport system components EscAB, and its ill-fitting catalytic motif structure.

Mca10 (*M.capsulatus str. Bath* 1,068aa, gi:53804058) and **Mca12** (839aa, gi:53804062) are encoded 867bp apart, codirectionally transcribed, the gene encoding Mca12 being upstream. As such only one expression unit and gene neighborhood will be described for them. Three hypothetical protein genes that overlap the Mca12 gene and each other leave only 42bp of intergenic space between the two FUPA31 genes. The hypothetical proteins are of 87-118aa in length and none of them have significant homology to any proteins of known function. Two proteins are encoded codirectionally downstream of Mca10; first, one 40bp away encoding a protein identified as a 1 TMS, 146aa FUPA31 fragment and then 46bp later, one encoding a protein involved in catabolism of external DNA, having possible methyltransferase activity. Downstream 38bp from these are two genes, 41bp apart, there are two genes transcribed convergently to the operon of interest, encoding NADH dehydrogenase (EC 1.6.99.3) (no known metal cofactors) and adenylate cyclase (EC 4.6.1.1) (Co^{2+}, Mg^{2+} and Mn^{2+} all greatly increase activity, Ca^{2+} inhibits it [Yang JK, et al., 1983]). Neither of these genes is found near the FUPA31 genes with high frequency however. Upstream of the Mca12 gene,

only one gene encoding a 1 TMS, 88aa hypothetical protein with no significant homology is transcribed codirectionally, 190bp away. Two genes are transcribed divergently from that one. The first is 347bp away and encodes a 58aa hypothetical protein with no significant homology, the second is 136bp later and encodes a Ca^{2+}/H^{+} exchanger (TC# 2.A.19.2.1). The next gene, which is 6bp away, is convergently transcribed and encodes cytochrome c peroxidase family protein. This gene in turn has a 4-gene operon transcribed divergently from it, starting 241bp away. The proteins encoded by this operon are (in order of transcription):

1) hypothetical protein: FIG00808725, which is 58% homologous to phosphate-selective porin O and P superfamily protein and contains GIM5 chaperone-type domain

2) 26bp later, a putative lipoprotein

3) 180bp later, putative FMN-binding domain-containing protein

4) -3bp later (overlap), 3TMS, 178aa hypothetical protein with no significant homology

The genes divergently transcribed from the FUPA31-encoding operon occur near the operon at high frequency, and the ones further downstream of them occur at high enough frequency to be of significance when considering the possible function of the FUPA31 proteins.

Mca11 (673aa, gi:53802306) is encoded codirectionally 1193bp downstream of Mca2 (TC# 3.A.3.2.21, Ca^{2+}PA, 1,031aa, gi:53802308). Between these genes and with 3bp overlap of the Mca2 gene is one encoding S-adenosylmethionine synthetase (EC 2.5.1.6) (divalent cations are required for tripolyphosphatase activity but Ca^{2+} cannot

substitute Mg^{2+} like Mn^{2+} can in this enzyme [Markham GD, et al., 1980]). This gene is commonly found with homologous Ca^{2+}PAs (See *Anabaena, Nostoc, Nitrosospira, Roseiflexus, Desulfobacca, Thermodesulfotator and Polymorphum*). After another 44bp downstream of the **Mca11** (FUPA31) gene, there are 3 more co-directional genes encoding first a HAMP domain-containing protein 46% homologous to the N-terminal half of multi-sensor signal transduction histidine kinase, then 61bp after that, ParA family protein, a putative cobyrinic acid a,c-diamide synthase and cell division inhibitor and finally 36bp after that a 100aa hypothetical protein in the HTH superfamily and thus a putative transcriptional regulator with 65% homology to the full, ~350aa proteins. Upstream of the Mca2 (Ca^{2+}PA) gene, 337bp away, a single gene is divergently transcribed which encodes 5'-nucleotidase (EC 3.1.3.5) (Mg^{2+}, Mn^{2+} and Co^{2+} can all activate the enzyme. Ni^{2+} may as well [Proudfoot M, et al., 2004] or it may inhibit it [Itami H, et al., 1989]). Convergently transcribed to that, 13bp away, is the last gene of 3-gene operon. The first two genes in this operon are predicted in SEED to encode homologues 47bp apart which are confirmed to be 29.2% homologous by GSAT binary alignment, but they are both 1TMS hypothetical proteins with no significant homology, of 356 and 347 aas in length respectively, and thus do not indicate a function for the FUPA31 protein encoded nearby. The third gene, 35bp away, encodes a Fis family sigma54-dependent transcriptional regulator that may be an anaerobic nitric oxide reductase transcriptional regulator, NorR. All of the other genes in the more distant gene neighborhood are found at such low frequency that they can be of little relevance, and most of them encode hypothetical proteins with no significant homology besides. There

are 2 genes, found 5.5kb and 8.7kb downstream of the Mca11 gene, which both encode diguanylate cyclase/phosphodiesterase (GGDEF & EAL domains) with PAS/PAC sensors of 464 and 594 aas in length respectively. These genes are found near the Mca11 gene at high frequency, in stark contrast to their neighbors.

Methylocella silvestris BL2 also encodes 2 P-type ATPases which have been identified by TCDB as FUPA31s. They will be referred to as **Msi12** (876aa, gi:217978977) and **Msi13** (909aa, gi:217978976) because they are both most similar to the *M.capsulatus* FUPA31 "Mca12". They are transcribed 622bp apart divergently, with a single 133aa hypothetical protein encoded in the same direction as Msi12, overlapping it by 3bp and 224bp away from the Msi13 gene. This likely transcription unit is flanked on both sides by convergently transcribed genes. The Msi12 gene is met 36bp away, convergently by a 3-gene operon encoding sulfate & thiosulfate binding protein CysP, then 64bp later sulfate permease and 338bp after that, TerC family integral membrane protein. The Msi13 gene is met 28bp away, convergently by a 3-gene operon encoding a 45aa hypothetical protein with no significant homology, then 135bp later exopolysaccharide synthesis protein ExoD and then with a 3bp overlap, a 106aa hypothetical protein with no significant homology. Transcribed divergently 424bp away from this operon are two genes encoding a putative amino acid permease followed 263bp later by a 149aa hypothetical protein with no significant homology. Homologues of these last two proteins are also encoded 6.7kb further downstream.

Beijerinckia indica subsp. indica ATCC 9039 also encodes 3 P-type ATPases which have been identified by TCDB as FUPA31s. They will be referred to as **Bin12** (872aa, gi:182679732), **Bin13** (925aa, gi:182678665) and **Bin 14** (889aa, gi:182678669) because they too are all most similar to the *M.capsulatus* FUPA31 "Mca12".

The **Bin12**-encoding gene is transcribed with two other co-directional genes upstream of it and another one transcribed divergently 138bp away from the first of those. The co-directional genes encode LysR family transcriptional regulator PA2877, followed 181bp later by a 131aa hypothetical protein with no significant homology which ends 75bp before the Bin12 gene. Convergently transcribed to the Bin12 gene is the 12th gene in an operon at least partially involved in inositol catabolism/utilization that is part of a bidirectional transcription unit that continues 7 genes in the opposite direction as well. The convergent operon encodes the following proteins in order of transcription:

1) 1 TMS, 50aa hypothetical protein with no significant homology

2) -28bp (overlap) later, 2-ketoglutaric semialdehyde dehydrogenase (EC 1.2.1.26) (no known metal requirements)

3) 12bp later, 2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase (EC 4.1.2.-)

4) 2bp later, D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95) (Mg^{2+} required [Peters-Wendisch, et al., 2005])

5) 251bp later, Myo-inositol 2-dehydrogenase (EC 1.1.1.18) (no known metal requirements)

6) 16bp later, Myo-inositol 2-dehydrogenase (EC 1.1.1.18) (duplicate)

7) 0bp later, Inososе isomerase (EC 5.3.99.-)

8) -3bp (overlap) later, Oxidoreductase (EC 1.1.1.-)

9) 137bp later, Inositol transport system sugar-binding protein

10) 58bp later, Inositol transport system ATP-binding protein

11) 46bp later, Inositol transport system permease protein

12) 163bp later, Dihydroxy-acid dehydratase (EC 4.2.1.9) (divalent metal ion required, Fe^{2+} most effective but Mg^{2+} and Mn^{2+} also work [Myers JW, 1961])

Divergently transcribed from this operon, 216bp away a 7-gene operon encodes the following proteins in order of transcription:

1) 5 TMS integral membrane protein; COG0730 superfamily, predicted permease

2) 247bp later, predicted transcriptional regulator of the myo-inositol catabolic operon

3) 114bp later, Epi-inositol hydrolase (EC 3.7.1.-)

4) 119 later, Inososе dehydratase (EC 4.2.1.44) (Co^{2+} and Mn^{2+} both increase activity [Berman T, et al., 1966])

5) -19bp later, Oxidoreductase, short chain dehydrogenase/reductase family

6) 314bp later, Xylose ABC transporter, periplasmic xylose-binding protein XylF

7) 317bp later, Carbohydrate-selective porin-like protein, OprB family

The gene divergently transcribed from the Bin12-encoding operon encodes short-chain dehydrogenase/reductase SDR. The gene transcribed convergently to it, 159bp away, is the third of a possible 3-gene operon encoding a 42aa hypothetical protein with

no significant homology, then 29bp later AsnC-family transcriptional regulator and finally a 188aa probable transmembrane protein of unknown function.

The **Bin13** and **Bin14** proteins are encoded 3471bp apart, codirectionally transcribed, with Bin13 encoded downstream. In the space between them, two copies of aspartate/ glutamate/ uridylate kinase are encoded in the opposite direction, 13bp apart. They are 211bp away and convergently transcribed from Bin14, 1602bp away and divergently transcribed from Bin13. The genetic landscape upstream of the Bin14 gene is of very little interest as none of the genes are found with appreciable frequency near FUPA31 genes, but those that may potentially compose a transcription unit involving the Bin14 gene will still be described. One co-directional gene is found upstream of the Bin14 gene, 231bp away, encoding ribonuclease I precursor (EC 3.1.27.6) (Cd^{2+}, Co^{2+}, Mg^{2+} and Mn^{2+} are all found to reverse EDTA- and o-phenanthroline-induced inhibition [Levy CC, et al., 1970]. Three small genes make up a putative operon that is transcribed divergently 110bp away from this that may be derived from a single fragmented gene. The first of these encodes a protein described as a putative LysR family transcriptional regulator, but considering these regulators are typically ~300aa in length and this one is only 69aa, it is unlikely to be a complete protein. The second and third genes in the operon are 285bp and 58bp later, respectively, and encode hypothetical proteins with a HdeA superfamily domains. The first has 1 TMS and is 100aa long, the second has no TMS and is 86aa long.

Downstream of the Bin13 gene, 6 genes are found to be in common with those surrounding the Msi12 and Msi13 -encoding transcription unit in *M.silvestris*. The first

two are part of a 4-gene putative operon which begins with the Bin13 gene and continues 189bp later with genes encoding first sulfate & thiosulfate binding protein CysP, then a 58aa hypothetical protein with no significant homology is encoded with a 25bp overlap and finally sulfate permease is encoded with an overlap of 3bp. There are four genes downstream of this, which are all transcribed in the convergent direction but which are separated into two bicistrons by an open 1,543bp intergenic space. The closer two genes encode exopolysaccharide synthesis protein ExoD followed 45bp later by a 95aa hypothetical protein, homologous to the 106aa one in *M.silvestris*, which ends 87bp from the Bin13 gene. The other pair encodes a putative amino acid permease followed 464bp later by a 121aa hypothetical protein homologous to those encoded following the two copies of the putative amino acid permease gene in *M.silvestris*. A TCDBLAST of these putative amino acid permeases reveals them to be members of the Amino Acid-Polyamine-Organocation (APC) Superfamily (TC 2.A.3.-.-) and CDD shows that they contain “spore germination” permease family domain. The remaining genes in this direction encode hypothetical proteins with no significant homology and thus will not be elaborated on

As mentioned above, no RegPredict analysis is currently available for the organisms encoding these FUPA31 proteins.

Even after expanding the sample size for this family, very limited patterns emerge to suggest possible functions of these P-type ATPases. The shared genes encoding ExoD & associated 100aa protein, Sulfate/thiosulfate binding protein CysP,

sulfate permease and amino acid permeases between two of the three examined species are the strongest evidence of shared function, so exopolysaccharide biosynthesis, cysteine biosynthesis, inorganic sulfur assimilation and organocation transport may all be candidate functions. Alternatively, taking the simpler approach of examining the metal requirements of the surrounding genes; divalent cation transport, especially of Mg^{2+} or Mn^{2+}, but also Co^{2+}, may be candidate functions as well.

The FUPA32 ATPase Family

FUPA32 is a family of Type I P-type ATPases (Chan, et al., 2010) with 8 subfamilies representing orthologues in (1) Proteobacteria, (2) Actinobacteria, (3) Firmicutes, (4) Fusobacteria, (5) Spirochaetes, (6) Euryarchaeota, (7) Verrucomicrobia and (8) Cyanobacteria. Those analyzed here range in size from 606-770 aas and have topologies of two TMS in subfamilies 1, 3, 6 and 8; four TMS in subfamilies 4 and 5; six TMS in subfamily 7 and eight TMS in subfamily 2 according to TCDB. As described previously, the WHAT and HMMTOP programs appears not to count numerous regions of high hydropathy and low amphipathicity in the low-TMS count proteins, suggesting that several other TMS in fact exist. The first two, which correspond to the standard TMS 1 & 2 notation cluster within a 41 aa range found between residues 39-141, 3 & 4 cluster within a 47 aa range about 15 residues away from 1 & 2. TMS 5 & 6 cluster together within a 45 aa range about 140 residues away from 3 & 4. TMS 7 & 8 cluster together within a range of 38 aas about 250-290 residues away from 5 & 6 and 5-45 residues from the end of the protein. The family's nearest hits in known-function families in TCDB are Family 5 (TC 3.A.3.5.-), subfamilies 15 and 18 and Family 6 (TC 3.A.3.6.-), subfamilies 13 and 17. However, in FUPA32 subfamilies 3, 5, 6, 7 and 8, FUPA31 comes up in a TCBLAST before Families 5 or 6.

TC 3.A.3.32.1

α-, β-, δ- and ε-Proteobacteria

Rpa7 (*Rhodopseudomonas palustris* CGA009 698aa, gi:39935402), which is frequently referred to as CtpC, is encoded as the last gene of a 6-gene operon. The operon encodes the following genes, in order of transcription:

1. 2 TMS, 166aa protein identified as a copper chaperone in SEED despite no copper chaperone hits in 3 iterations of PSIBLAST. The protein does show 35% similarity to the N-terminus of FUPA32 (TC# 3.A.3.32.2), including the region containing an HMA domain, and 46% similar homologues are found to possess this domain in CDD, however, this protein does not.
2. 8bp later, 110aa hypothetical protein referred to as FIG01006922 with 40% similarity to N-terminus of FUPA32 (TC #3.A.3.32.5), possessing a HMA domain
3. 18bp later, 1TMS, 107aa hypothetical protein: FIG01005902 with no significant homology
4. -13bp (overlap) later, 113aa hypothetical protein with no significant homology
5. -7bp (overlap) later, putative oxidoreductase
6. 3bp (overlap) later, FUPA32

Convergently transcribed to this operon, 236bp away, is an 11-gene operon which includes at least 5 genes probably involved in utilization of glutathione as a sulfur source. The operon encodes the following proteins in order of transcription:

1. LysR-family transcriptional regulator
2. 168bp later, gamma-aminobutyrate:alpha-ketoglutarate aminotransferase (EC 2.6.1.19) (no iron-sulfur cluster [Liu W, et al., 2004])

3. 38bp later, succinate-semialdehyde dehydrogenase [NADP+] (EC 1.2.1.16) (bivalent cations such as Mg^{2+}, Mn^{2+}, Ca^{2+} or Fe^{2+} are not required [Sanchez M, et al., 1989])
4. 50bp later, acetylornithine deacetylase (EC 3.5.1.16) (addition of Co^{2+} creates an 8x fold increase in activity, Zn^{2+} creates a 2x fold increase at 0.1mM but inhibits activity at higher concentrations, PO_4^{3-} also activates [Javid-Majd F, et al., 2000])
5. 0bp later, COG0028: thiamine pyrophosphate-requiring enzyme: putative acetolactate synthase large subunit/pyruvate decarboxylase

   79bp later, putative glutathione ABC transporter, solute-binding component
6. -3bp (overlap) later, putative glutathione ABC transporter, permease component
7. 13bp later, putative glutathione ABC transporter, permease component
8. -3bp (overlap) later, putative glutathione ABC transporter, ATP-binding component
9. 175bp later, gamma-glutamyltranspeptidase (EC 2.3.2.2) (1mM Na^+, K^+, Mg^{2+} or Mn^{2+} create >28% stimulation, 1mM Cu^{2+}, Hg^{2+}, Ni^{2+} or Zn^{2+} create strong inhibition [Yao YF, et al., 2006])
10. 294bp later, nuclease (SNase domain protein)

Divergently transcribed from the FUPA32-encoding operon, 498bp away, is a single gene encoding iron-responsive regulator Irr. This gene in turn has a 3-gene operon transcribed convergently to it, 127bp away. This operon first encodes a major facilitator superfamily protein of the metabolite:H^+ symport subfamily (TC# 2.A.1.6.-), probably a member of subfamily 5 or 6, transporting 4-methyl-o-phthalate or shikimate,

respectively, as TCBLAST scores for these families are exceptionally higher than the rest, both at e-value=e-72. This is followed 65bp later by dihydroxy-acid dehydratase (EC 4.2.1.9) (divalent metal ion required, Fe^{2+} is most effective, Mg^{2+} and Mn^{2+} also work [Myers JW, 1961]) which is followed 52bp later by FAD dependent oxidoreductase.

Aar5 (*Aromatoleum aromaticum* EbN1 694aa, gi:56475751) is encoded ninth gene in a putative 11-gene operon which encodes the following proteins in order of transcription:

1. FIG135464: Cytochrome c4
2. 11bp later, FIG002261: Cytochrome c family protein, putative c5
3. 157bp later, 2 TMS putative copper chaperone with 44% similar to N-terminus of FUPA32 (TC# 3.A.3.32.1)
4. 27bp later, 111aa hypothetical protein with 46% similar to N-terminus of FUPA32 (TC #3.A.3.32.1), possessing a HMA domain
5. -7bp (overlap) later, ferritin/ribonucleotide reductase-like protein
6. 2bp later, 1 TMS, 120aa hypothetical protein
7. -3bp (overlap) later, 2 TMS, 114aa hypothetical protein
8. -19bp later, putative oxidoreductase
9. -3bp (overlap) later, FUPA32
10. 138bp later, PA-phosphatase related phosphoesterase

11. -3bp (overlap) later, diacylglycerol kinase (EC 2.7.1.107) ("enzyme requires a free divalent metal cation: Mg^{2+}, Mn^{2+}, Co^{2+}, Cd^{2+} or Zn^{2+}" [Walsh JP, et al., 1992])

The 2 proteins homologous to FUPA32 N-terminus without and with a HMA domain and the putative oxidoreductase encoded just before the FUPA32 gene here and in the Rpa7-encoding operon described above are respectively homologous. Genes 6-8 listed above may actually be a single gene, *ligA* encoding a protein called LigA.

Convergently transcribed to this operon, 100bp away, a single gene is transcribed which encodes tryptophanase (EC 4.1.99.1) (Mg^{2+} binds to enzyme, Cl^- bound to enzyme, may be required fot stabilization of subunit interactions, essentially any monovalent cation required for activity: Li^+, Na^+, K^+, Rb^+, Cs^+, NH_4^+ and Tl^+ [Suelter CH, et al., 1977; Tsesin N, et al., 2007])

Divergently transcribed from the FUPA32-encoding operon, 446bp away, a 2-gene operon is found encoding first small-conductance mechanosensitive channel followed 196bp later by prophage Lp2 protein 6. This operon is in turn met convergently, 637bp later, by a possible 7-gene operon encoding the following proteins in order of transcription:

1. aspartokinase (EC 2.7.2.4) (requires Mg^{2+} or Mn^{2+} [Dungan SM, et al., 1973])
2. 61bp later, tRNA-Ser-GCT
3. 160bp later, 77aa hypothetical protein with no significant homology
4. 190bp later, 519aa mobile element protein, integrase catalytic subunit

5. 23bp later, 244aa mobile element protein, putative transposase ATP-binding subunit
6. 375bp later, 76aa hypothetical protein with no significant homology
7. 31bp later, 452aa mobile element protein, IS4 family transposase

Dvu4 (*D.vulgaris* str. Hildenborough 606aa, gi:46581732) is encoded as the third gene of a putative 4-gene operon which first encodes a 1 TMS, 91aa hypothetically iron-regulated protein referred to as FIG00607321 with no significant homology, followed 76bp later by 127aa hypothetical protein, also with no significant homology. The FUPA32 gene overlaps this by 9bp and is itself overlapped by 3bp by a gene encoding another homologous 2 TMS putative copper chaperone with 42% homology to the N-terminus of FUPA32 but lacking an HMA domain. This operon is met convergently, 293bp away, by a 6-gene operon encoding the following proteins in order of transcription:

1. Potassium-transporting ATPase (PTA) A chain KdpA (EC 3.6.3.12) (TC 3.A.3.7.1)
2. 17bp later, PTA B chain KdpB, Dvu5 (678aa, gi:46581738 [Chan, et al., 2010]) (EC 3.6.3.12) (TC 3.A.3.7.1)
3. 10bp later, PTA C chain KdpC (EC 3.6.3.12) (TC 3.A.3.7.1)
4. 82bp later, osmosensitive K^+ channel histidine kinase KdpD (EC 2.7.3.-)
5. -3bp (overlap) later, sensory box histidine kinase

6. -3bp (overlap) later, response regulator of zinc sigma-54-dependent two-component system

Divergently transcribed from the FUPA32-encoding operon, 45bp away, a putative 4-gene operon encoding two hypothetical proteins, 8bp apart, the first 41aa long and the second 109aa long and containing 1 TMS. The next 2 genes are 322bp away and encode spermidine export proteins MdtJI, 9bp apart.

Wsu4 (*Wolinella succinogenes* DSM 1740 716aa, gi:34557896) is encoded as the twenty-first of a 22-gene putative operon, although 4 of the 5 genes transcribed right before it are the only ones predicted by SEED to be frequently transcribed with it. The full putative operon encodes the following proteins, in order of transcription:

1. flagellar P-ring protein FlgI
2. 12bp later, putative flagellar rod assembly muramidase FlgJ
3. -3bp (overlap) later, flagellar sensory histidine kinase FlgS
4. 98bp later, non-specific DNA-binding protein Dps / iron-binding ferritin-like antioxidant protein / ferroxidase (EC 1.16.3.1) (requires Fe, inhibited by Zn and terbium (Tb) [Havukainen H, et al., 2008])
5. 130bp later, magnesium and cobalt efflux protein CorC
6. -25bp (overlap) later, histidinol dehydrogenase (EC 1.1.1.23) (Mn^{2+} and Zn^{2+} at 0.5 mM enhance activity by 165% and 20%, respectively [Andorn N, et al., 1982])

7. 84bp later, superoxide dismutase [Fe] (EC 1.15.1.1) (may bind $1Cu^{+}$ and $1Zn^{+}$ each [Beyer,W et al., 1991])
8. 10bp later, 86aa hypothetical protein: DUF2018 superfamily -expressed with Fe associated genes
9. 3bp later, 1 TMS, 146aa hypothetical protein
10. -22bp (overlap) later, octaprenyl-diphosphate synthase (EC 2.5.1.-) / dimethylallyltransferase (EC 2.5.1.1) (*may* require Mg^{2+} [Schmidt A, et al., 2010])/ geranyltranstransferase (farnesyldiphosphate synthase) (EC 2.5.1.10) (absolute dependence on presence of divalent cation, Mg^{2+} allows optimal activity but Fe^{2+} and Mn^{2+} also work [Dhiman RK, et al., 2004])/ geranylgeranyl pyrophosphate synthetase (EC 2.5.1.29) (Mg^{2+} required for activity, Mn^{2+} also activates slightly [Sagami H, Ogura K, 1981,9185])
11. 2bp later, glutamyl-tRNA reductase (EC 1.2.1.70) (Mg^{2+} stimulates activity and restores it after treatment with chelating agents, Ca^{2+} and Mn^{2+} also restore activity [Schauer S, et al., 2002])
12. 8bp later, prolyl-tRNA synthetase (EC 6.1.1.15) (requires Mg^{2+} [Crepin T, et al., 2006])
13. 0bp later, FxsA cytoplasmic membrane protein
14. -3bp (overlap) later, porphobilinogen deaminase (EC 2.5.1.61) (no known metal cofactor requirements)
15. 19bp later, putative outer membrane porin

16. 162bp later, 2 TMS putative copper chaperone with 53% similarity to N-terminus of FUPA32 (TC# 3.A.3.32.3)
17. -13bp (overlap) later, putative ferritin
18. 4bp later, 1 TMS, 105aa hypothetical protein with no significant homology
19. -3bp later, 1TMS, 108aa hypothetical protein with 58% similarity to Cys/Met metabolism pyridoxal-phosphate-dependent enzyme
20. 21bp later, putative oxidoreductase/RNA methyltransferase
21. -22bp (overlap) later, FUPA32
22. 109bp later, 2 TMS, 73aa hypothetical protein with 67% similarity to putative helicase2C

Convergently transcribed to this operon, and overlapping it by 3bp, is a bicistron encoding first a putative iron permease FTR1 fragment about one third the size of the full protein, followed, with a 3bp overlap, by HAMP + GGDEF(DGC) domain protein (Histidine kinase, Adenylyl cyclase, Methyl-accepting protein, and Phosphatase + DiGuanylate Cyclase). Transcribed divergently from this operon in turn, 172bp away, is an 8-gene operon encoding the following proteins in order of transcription:

1. putative high-affinity iron permease
2. 55bp later, periplasmic protein p19 involved in high-affinity Fe^{2+} transport
3. 9bp later, 7 TMS integral membrane protein: putative tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA (Glucose-inhibited division protein A)
4. -3bp (overlap) later, cell division protein FtsX

5. -13bp (overlap) later, cell division protein FtsX (duplicate)
6. 5bp later, ABC transporter, ATP-binding protein, member of LPT/MacB family
7. -3bp (overlap) later, putative thioredoxin precursor
8. -3bp (overlap) later, putative lipoprotein thioredoxin

Divergently transcribed from the FUPA32-encoding operon, 142bp away, is a single gene encoding tRNA-Pro-GGG.

Convergently transcribed to the tRNA gene, 34bp away, is a 3-gene operon encoding agmatine deiminase (EC 3.5.3.12) (no known metal cofactor requirements) followed 3bp later by alpha-aspartyl dipeptidase peptidase E (EC 3.4.13.21) (the enzyme contains no metals [Conlin CA, et al., 1994]) and 367bp after that, a 90aa hypothetical protein with no significant homology.

Divergently transcribed from this tricistron is another, beginning 84bp away, which encodes hydrogenase-4 component G, followed 16bp later by YgjD/Kae1/Qri7 family, required for threonylcarbamoyladenosine (t(6)A) formation in tRNA, which is followed 56bp later by thiol peroxidase, Tpx-type (EC 1.11.1.15) (no known metal ion requirements).

RegPredict analysis of co-regulation with FUPA32 genes in Proteobacteria was done using intergenic sequence preceding the predicted first gene in each FUPA32-encoding operon. In *R.palustris*, this is the gene encoding 2 TMS putative copper chaperone with 35% similar to N-terminus of FUPA32, in *A.aromaticum* it is Cytochrome c4 and 2 TMS putative copper chaperone (2TMSpCuCh) with 44% similar

to N-terminus of FUPA32, in *D.vulgaris* it is 1 TMS, 91aa hypothetical protein and in *W.succinogenes* it is Flagellar P-ring protein FlgI, Magnesium and cobalt efflux protein CorC and 2 TMS putative copper chaperone with 53% similar to N-terminus of FUPA32.

The first predicted sequence motif is derived from predicted regulatory sequences in all four organisms, preceding The FUPA32 gene and a few other upstream genes in the respective operons. These sequences, as well as their upstream of the nearest downstream gene in the operons are described below:

A.aromaticum	2TMSpCuCh	119bp	CGgaTgGcG|CtCgAagCG, score=4.57
R.palustris	2TMSpCuCh	103bp	GcCaGtgCG|CGatCcGcC, score=5.61
	put. oxidored.	85bp	GGCGggGcG|CtCggCGCC, score=4.56
D.vulgaris	FUPA32	115bp	ggCgGgGCG|CGCaCaGgg, score=4.46
W.succinogenes	CorC	105bp	GagaTGGcG|CtCCActcC, score=4.79

The first member of this motif set is a sequence (GCCGGgGcG|CtCtCCGGC, score=4.86) found only in *A.aromaticum*, 71bp upstream of an operon encoding the following proteins:

1. Acyl-CoA dehydrogenase, long-chain specific, mitochondrial precursor (EC 1.3.99.13) (no known metal requirements)
2. Acyl-CoA dehydrogenase, short-chain specific (EC 1.3.99.2) (no known metal requirements)
3. Inosine-5'-monophosphate dehydrogenase (EC 1.1.1.205) (no known divalent cation requirements)

4. *korC1* encodes Glutamate synthase [NADPH] small chain (EC 1.4.1.13) (iron-sulfur protein [Miller RE, et al., 1972])
5. *korA1* encodes 2-oxoglutarate oxidoreductase, alpha subunit (EC 1.2.7.3) (2[4Fe-4S] cluster per ferredoxin-type enzyme [Dörner E, et al., 2002])
6. *korB1* encodes 2-oxoglutarate oxidoreductase, beta subunit (EC 1.2.7.3)

The second member of this set is also found only in *A.aromaticum*. The sequence (gcCagggCG|CGaaaaGcg, score=4.92) is 107bp upstream of a bicistron encoding Fumarate reductase/succinate dehydrogenase flavoprotein, N- terminal and Polyvinyl-alcohol dehydrogenase, PQQ-dependent (gene: *pvaA*) (EC 1.1.99.23) (no known metal requirements)

The third member is found only in *R.palustris*, as are the rest of the sequences in this set. This sequence (gcCggTGCG|CGCAaaGcg, score=4.81) is found 129bp upstream of a monocistron encoding Cytochrome c family protein.

The fourth member of this motif set is a sequence (GagaAtGCg|gGCtTctcC, score=4.75) found 172bp upstream of a 4-gene operon consisting of genes *bchCXYZ* which encode 2-desacetyl-2-hydroxyethyl bacteriochlorophyllide A dehydrogenase BchC followed by Chlorophyllide reductase subunits BchXYZ (EC 1.18.-.-).

The fifth member of this motif set is a sequence (GcgaggGCG|CGCttggcC, score=4.78) found 238bp upstream of a tricistron encoding the following proteins:

1. 16S rRNA m(5)C 967 methyltransferase (EC 2.1.1.-)
2. Heparinase II/III-like

3. *purH* encodes IMP cyclohydrolase (EC 3.5.4.10) (requires Mg^{2+} [Kang YN, et al., 2007])/ Phosphoribosylaminoimidazolecarboxamide formyltransferase (EC 2.1.2.3) (no known divalent cation requirements)

The second predicted sequence motif is derived from predicted regulatory sequences in 2 genomes, preceding The FUPA32 gene and a few other upstream genes in the respective operons. These sequences, as well as their upstream of the nearest downstream gene in the operons are described below:

A.aromaticum	*cytC4*	21bp	CGgaaCtGC|GCgGaggCG, score=4.88
	ferritin gene	44bp	CGGaaCGcc|ttCGccCCG, score=4.90
R.palustris	2TMSpCuCh	65bp	caaaACGcC|GtCGTgccg, score=5.41

The first member of this motif set is a sequence (CaGaccgGC|GCgatgCcG, score=4.62) found in *R.palustris*, 103bp upstream of a tricistron encoding the following proteins:

1. Alpha/beta hydrolase fold (EC 3.8.1.5) (no known metal requirements)
2. *crtI* encodes Phytoene desaturase, neurosporene or lycopene producing (EC 1.3.-.-)
3. *crtB* encodes Phytoene synthase (EC 2.5.1.32) (Mn^{2+} or Mg^{2+} required [Neudert U, et al., 1998])

The second member of this motif is a sequence (CaGgACGgC|GaCGTgCcG, score=4.63) found in *A.aromaticum*, 135bp upstream of a tricistron encoding the following proteins:

1. *galE1* encodes UDP-glucose 4-epimerase (EC 5.1.3.2) (contains none of the common metal ions [Arabshahi A, et al., 1988])

2. *wbiH* encodes Glycosyl transferase, family 4 precursor
3. Nucleoside-diphosphate sugar epimerase/dehydratase

The third member of this motif is a sequence (CaggtCtcC|GtcGtggcG, score=4.70) found in *A.aromaticum*, 133bp upstream of a tricistron encoding the following proteins:

1. *msrA* encodes Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11) (Mg^{2+} required [Brot N, et al., 1981])
2. *fkpA* encodes Peptidylprolyl isomerase (EC 5.2.1.8) (Mg^{2+} suggested in Eukarya, but no data for prokaryotes [Jordens J, et al., 2005])
3. Cobalt-zinc-cadmium resistance protein

The third member of this motif is a sequence (CGaGCCGgC|GtCGGCcCG, score=4.67) found in *R.palustris*, 158bp upstream of a 6-gene operon containing *badDEFGAB*, which encode Benzoyl-CoA reductase subunits BadDEFG (EC 1.3.99.15) (no known metal requirements), Benzoate-CoA ligase BadA (EC 6.2.1.25) (Mg^{2+} required, Mn^{2+} can replace it [Geissler JF, et al., 1988]) and 4Fe-4S ferredoxin, iron-sulfur binding BadB.

The fourth member of this motif is a sequence (CAGgaCGgC|GtCGatCTG, score=4.62) found in *A.aromaticum*, 71bp upstream of a tricistron encoding the following proteins:

1. *iscU* encodes Iron-sulfur cluster assembly scaffold protein IscU
2. *iscA* encodes Iron binding protein IscA for iron-sulfur cluster assembly
3. *hscB* encodes Chaperone protein HscB

4. *hscA* encodes Chaperone protein HscA
5. *fdx* encodes Ferredoxin, 2Fe-2S
6. protein putatively involved in assembly of Fe-S clusters
7. *btuE* encodes Glutathione peroxidase (EC 1.11.1.9) (known metal requirements in prokaryotes)

The fifth member of this motif is a sequence (CaagcCtcC|GctGcgccG, score=4.68) found in *R.palustris*, 38bp upstream of a tricistron encoding NUDIX hydrolase, Potassium efflux system KefA protein / Small-conductance mechanosensitive channel and GCN5-related N-acetyltransferase.

The sixth member of this motif is a sequence (CGGgACGgC|GgCGTgCCG, score=4.63) found in *R.palustris*, 91bp upstream of a tricistron encoding possible L-sorbosone dehydrogenase, Cytochrome c4 and possible Acyl-CoA dehydrogenase.

The third predicted sequence motif is derived from predicted regulatory sequences in all four organisms, preceding The FUPA32 gene and a few other upstream genes in the respective operons. These sequences, as well as their upstream of the nearest downstream gene in the operons are described below:

A.aromaticum	*cytC4*	97bp	CGCgGCtat|gctGCgGCG, score=5.16
R.palustris	2TMSpCuCh	60bp	CGcCGtcGt|gCcgCGcCG, score=5.27
D.vulgaris	1 TMS, 91aa hyp.	212bp	CTGtGccAt|cTccCgCAG, score=4.94
W.succinogenes	His-dehydrogenase	179bp	cGcgggtgt|ggtgagcCt, score=4.4

The first member of this motif set is a sequence (CGctGcaat|gaccCgcCG,

score=4.85) found only in *R.palustris*, 81bp upstream of an operon encoding the following proteins:

1. 2 TMS, 86aa hypothetical protein with no significant homology
2. transcriptional regulator, Crp/Fnr family
3. 1 TMS, 62aa hypothetical protein in DUF2892 superfamily
4. Possible carboxymuconolactone decarboxylase family protein (EC 4.1.1.44) (no known metal requirements)
5. *ribB* encodes 3,4-dihydroxy-2-butanone 4-phosphate synthase / GTP cyclohydrolase II (EC 3.5.4.25) (requires either Mg^{2+} or Zn^{2+}, both found natively [Kaiser J, et al., 2002; Blau N, et al., 1985])
6. Glutaryl-CoA dehydrogenase (EC 1.3.99.7) (no known metal requirements)
7. Metallo-beta-lactamase family protein

The second member of this motif set is a sequence (CgGCGgcat|cccgCGCaG, score=4.74) found only in *D.vulgaris*, 58bp upstream of a 4-gene operon encoding the following proteins:

1. phosphoribosylaminoimidazolecarboxamide formyltransferase, putative
2. Exoribonuclease II (EC 3.1.13.1) (Mg^{2+} and K^{+} are required for activity [Bollenbach TJ, et al., 2004])
3. Acyl-phosphate:glycerol-3-phosphate O-acyltransferase PlsY
4. *thrC* encodes Threonine synthase (EC 4.2.3.1) (no known metal requirements)

The third member of this motif set is a sequence (CGcTGccAa|gTccCAcCG, score=4.62) found only in *R.palustris*, 50bp upstream of an operon encoding the following proteins:

1. possible photosynthetic complex assembly protein, puhC
2. 99aa hypothetical protein with no significant homology
3. Mg protoporphyrin IX monomethyl ester oxidative cyclase (aerobic) (EC 1.14.13.81) (Mg^{2+} required [Gough SP, et al., 2007])
4. photosynthetic complex assembly protein 2, puhE (7TMS integral membrane protein)
5. *hemA* encodes 5-aminolevulinate synthase (EC 2.3.1.37) (no known metal requirements in prokaryotes)

The fourth member of this motif set is a sequence (CGcgGgtGC|GCggCgcCG, score=4.97) found only in *A.aromaticum*, 91bp upstream of a tricistron encoding the following proteins:

1. Inactive homolog of metal-dependent proteases, putative molecular chaperone
2. Ribosomal-protein-S18p-alanine acetyltransferase (EC 2.3.1.-)
3. Uracil-DNA glycosylase, family 4

The fifth member of this motif set is a sequence (cGcGGCCGt|gCGGCCcCc, score=4.99) found only in *R.palustris*, as are the remaining members of this set, 150bp upstream of a bicistron encoding transcriptional regulator, Crp/Fnr family and putative diguanylate cyclase (GGDEF)/phosphodiesterase (EAL) with PAS domain.

The sixth member of this motif set is a sequence (CGCCGCcGa|cCcGCGGCG, score=4.81) found 49bp upstream of a bicistron encoding putative cytochrome P450 hydroxylase and putative acyl-CoA dehydrogenase.

The seventh and last member of this motif set is a sequence (CGtCGtcGt|gCtgCGgCG, score=4.96) found 196bp upstream of a monocistron encoding probable AraC family transcriptional regulator.

The fourth predicted sequence motif is derived from predicted regulatory sequences in 3 genomes, preceding The FUPA32 gene and a few other upstream genes in the respective operons. These sequences, as well as their upstream of the nearest downstream gene in the operons are described below:

A.aromaticum	*cytC4*	100bp	CGCCGCgGCt|A|tGCtGCGGCG, score=4.66
	ligA	165bp	CGtcGtcGgc|C|cgCtcCtgCG, score=4.17
R.palustris	2TMSpCuCh	60bp	cGcCGtCGtG|C|CgCGcCGcCc, score=5.05
W.succinogenes	*flgI*	187bp	CaCcCACCCc|A|aGGGTGtGcG, score=4.47
		189bp	CtCACCCacc|C|caaGGGTGtG, score=4.63

The first member of this motif set is a sequence (cGCgcCcGCg|A|tGCtGcgGCc, score=4.66) found only in *R.palustris*, 148bp upstream of a bicistron consisting of *msrA1B*, which encode methionine sulfoxide reductase A1 and Peptide methionine sulfoxide reductase MsrB (EC 1.8.4.12) (binding sites for both Fe^{2+} and Zn^{2+} [Olry A, et al., 2005])

The second member of this motif set is a sequence (cGCCGagatc|A|agcggCGGCt, score=4.71) found only in *A.aromaticum*, 86bp upstream of a monocistron which encodes

Aconitate hydratase (EC 4.2.1.3) (binds 1 4Fe-4S cluster per subunit [Tang Y, et al., 2005] and possibly Mg^{2+} [Tsuchiya D, et al., 2008]).

The third member of this motif set is a sequence (cGCCccCgCg|C|aGtGccGGCc, score=4.85) found only in *A.aromaticum*, 122bp upstream of a tricistron which encodes the following proteins:

1. Enoyl-CoA hydratase (EC 4.2.1.17) (Fe^{2+} in *Clostridium aminobutyricum*, each 56 kDa subunit of the homotetrameric enzyme contains one FAD and a [4Fe-4S]$^{2+}$ cluster, [Friedrich P,et al., 2008]) / 3,2-trans-enoyl-CoA isomerase (EC 5.3.3.8) (no known metal requirements in prokaryotes)/ 3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) (no known metal requirements)
2. Acetyl-CoA acetyltransferase (EC 2.3.1.9) (Mg^{2+} required, Mn^{2+} and Ca^{2+} can substitute to 90% activity [Kim SA, et al., 1997])
3. Enoyl-CoA hydratase (EC 4.2.1.17)

The fourth member of this motif set is a sequence (cGGCctCGCG|C|CGCGccGCCt, score=4.79) found only in *R.palustris*, 49bp upstream of an operon encoding the following proteins:

1. fds*A* encodes NAD-dependent formate dehydrogenase alpha subunit

2. *fdsC* encodes Formate dehydrogenase chain D (EC 1.2.1.2) (tungsten and iron are both found in this enzyme in *Methylobacterium extorquens* [Laukel M, et al., 2003] but "no divalent metal cation is needed as a cofactor for enzyme activity" in *Gelatoporia subvermispora*, a eukaryote [Watanabe T, et al., 2008])
3. *fdsD* encodes NAD-dependent formate dehydrogenase delta subunit
4. putative oxalate/formate Major Facilitator Superfamily (MFS) antiporter
5. possible L-sorbosone dehydrogenase
6. *cycC4* encodes Cytochrome c4
7. possible Acyl-CoA dehydrogenase
8. MoxR-like AAA ATPase
9. von Willebrand factor type A domain containing CoxE family protein
10. Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3) (Mg^{2+} required [Abe T, et al., 2008])

The fifth member of this motif set and last for the Proteobacterial group, is a sequence (cTcgcCacTG|C|CAagGagcAt, score=4.52) found only in *W.succinogenes*, 162bp upstream of an operon encoding the following proteins:

1. *aroQ* encodes 3-dehydroquinate dehydratase II AroC II (EC 4.2.1.10) (no known divalent ion requirements)
2. *pepQ* encodes proline aminopeptidase
3. *folK* encodes 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase (EC 2.7.6.3) (Mg^{2+} required, Ca^{2+} and Mn^{2+} can substitute [Ballantine SP, et al., 1994])

4. *flhF* encodes Flagellar biosynthesis protein FlhF
5. *fleN* encodes flagellar synthesis regulator FleN
6. Motility integral membrane protein
7. *fliA* encodes RNA polymerase sigma factor for flagellar operon
8. *fliM* encodes Flagellar motor switch protein FliM
9. *fliY* encodes Flagellar motor switch protein FliY
10. *trmU* encodes tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase (EC 2.1.1.61) (no known divalent ion requirements)
11. *prsA* encodes Ribose-phosphate pyrophosphokinase (EC 2.7.6.1) (Mg^{2+} required, Mn^{2+} equally effective, Zn^{2+} 25% effective, Co^{2+} and Ca^{2+} not effective [Switzer RL, 1969])
12. cyclophilin-type peptidyl-prolyl isomerase

TC 3.A.3.32.4

Fusobacteria

The three FUPA32s described in Chan, et al. (2010) are encoded in very similar operons and thus will not be described here entirely separately. The operon encoding Fnu8 is described first, then differences in the operons encoding the other two FUPA32s, Fnu9 and Fnu10, are noted.

Fnu8 (*Fusobacterium nucleatum subsp. nucleatum* ATCC 25586 735aa, gi:19704525) is encoded as the 5 gene in a 9-gene operon which encodes the following proteins in order of transcription:

1. N-acyl-L-amino acid amidohydrolase (EC 3.5.1.14) (Zn^{2+} or Co^{2+} required, Mn^{2+} and Ni^{2+} may substitute, Ca^{2+} and Mg^{2+} may not [Koreishi M, et al., 2009; Curley P, et al., 2003])
2. 47bp later, amino acid ABC transporter substrate-binding protein 240aa (TC# 3.A.1.3.15)
3. 179bp later, 3 TMS, 165aa hypothetical protein with 41% similarity to keratin, type II cytoskeletal I
4. 43bp later, 116aa hypothetical protein 50% similar (over 106aa) to the N-terminus of FUPA32 (TC #3.A.3.32.5)
5. 7bp later, FUPA32
6. 34bp later, 243aa conserved unknown protein only found downstream and codirectional to FUPA32 genes. It has no 45.8% similarity (e-value=0.0013) to voltage-gated Ca^{2+} channels according to TCDBLAST and 35% similarity to FUPA32 (TC #3.A.3.32.1) (e-value=5e-04)
7. 30bp later, 1 TMS, 90aa hypothetical protein 59% similar to Tat pathway signal sequence domain protein (e-value=7e-17)
8. 101bp later, 89aa hypothetical protein with no significant homology
9. 61bp later, 1 TMS, 68aa hypothetical protein

   Convergently transcribed to this is a 4-gene operon which ends with a 32bp

overlap of the aforementioned operon. This 4-gene operon encodes the following proteins, in order of transcription:

1. putative V-type ATPase transporter/abortive infection family protein RloA
2. -3bp (overlap) later, 213aa hypothetical protein (FIG00668306) abortive phage resistance protein, RloB superfamily
3. -10bp (overlap) later, 1 TMS, 177aa hypothetical protein referred to as FIG00670486
4. 273bp later, 211aa hypothetical protein with no significant homology

This operon in turn has a single gene transcribed divergently from it, 170bp away, encoding DNA polymerase IV (EC 2.7.7.7) (the holoenzyme requires Mg^{2+} and also Zn^{2+} [Kornberg T, et al., 1974 & Setlow P, 1974]).

Divergently transcribed from the FUPA32-encoding operon, 485bp away, a putative operon consisting of 10 co-directional genes is found. The proteins encoded in this operon, in order of transcription, are as follows:

1. NAD-dependent protein deacetylase of SIR2 family
2. 456bp later, abortive infection bacteriophage resistance protein AbiD
3. 28bp later, Clustered Regularly Interspaced Short Palindromic Repeats CRISPR-associated protein, TM1814 family
4. -10bp (overlap) later, CRISPR-associated protein Csx8 (FIG00669732)
5. 13bp later, CRISPR-associated negative autoregulator
6. 382bp later, short CRISPR-associated protein Cas5 (FIG00670073)
7. 89bp later, CRISPR-associated helicase Cas3

8. 43bp later, CRISPR-associated RecB family exonuclease Cas4a
9. -12bp (overlap) later, CRISPR-associated protein Cas1
10. -37bp (overlap) later, CRISPR-associated protein Cas2

The next gene downstream of this is 1.8kb away and convergently transcribed and thus will not be discussed as it is not likely to have bearing on the probable function of the FUPA32. It is worth noting that the RloB protein has strong hits in NCBI PSIBLAST as a CRISPR-associated protein as well and thus it is likely that the entire operon encoding the FUPA32 may have been inserted amongst a pre-existing CRISPR-associated/bacteriophage resistance protein encoding operon.

The operon encoding **Fnu9** (*F.nucleatum* subsp. *polymorphum* ATCC 10953 735aa, gi:167008397) is identical to that encoding Fnu8 except that it is missing the last gene, that encoding a hypothetical protein with 1 TMS and 168aa. The 4-gene operon encoding RloAB and two hypothetical proteins is also absent from this operon, which makes the gene encoding DNA polymerase IV part of the FUPA32-encoding operon, transcribed 31bp after the 89aa hypothetical protein gene.

The operon transcribed divergently to the FUPA32-encoding operon in *F.nucleatum subsp. polymorphum* is also very similar to that in *Fusobacterium nucleatum subsp. nucleatum*, with the exception that preceding the aforementioned divergent operon, there are an additional 7 genes transcribed codirectionally to it. The first of these begins 500bp from that encoding N-acyl-L-amino acid amidohydrolase. This additional 7-gene operon encodes Nickel ABC transporter components NikABCDE

(TC# 3.A.1.5.3), but *nikB* and *nikE* appear to be fragmented. A brief review of this operon's closest homologues in SEED suggests that this may be a sequencing error as it is the only case of fragmentation observed. Additionally, only *nikDE* appear to be found near the FUPA32 gene according to SEED, and this appears to be due to the high sequence homology among ATP-binding units of the ABC transporters, rather than any actual relation between these specific genes.

Fnu10 (*F.nucleatum subsp. vincentii* ATCC 49256 735aa, gi:34764203) is found only in an operon for the partially sequenced genome of *F.nucleatum subsp. vincentii*. As such, only the two genes transcribed immediately upstream of the FUPA32 gene can be confirmed the same for this genome, as the rest of the gene neighborhood is currently unknown.

Fusobacteria are not currently available for analysis in RegPredict.

TC 3.A.3.32.5

Tde1 (*Treponema denticola* ATCC 35405 699aa 42525989) is encoded as the second gene in a tricistron which also encodes putative Sec-independent protein translocase 27bp upstream and another homologous 2 TMS copper-chaperone-type protein with 42% similarity to the N-terminus of FUPA32, 20bp downstream.

Convergently transcribed to this operon, 49bp away, another tricistron is found, which encodes first spermidine/putrescine-binding protein followed 56bp later by Ni^{2+}-binding GTPase involved in regulation of expression and maturation of hydrogenase and then with a 10bp overlap, ABC transporter ATP-binding protein. A single gene encoding

Hcp transcriptional regulator HcpR (Crp/Fnr family) is divergently transcribed from this incompletely characterized ABC transporter cluster, 56bp away.

Divergently transcribed from the FUPA32-encoding operon, 138bp away, is a single gene encoding sialidase (EC 3.2.1.18) (Ca^{2+} and Ba^{2+} stimulate activity [Vertiev YV, et al., 1981]). This gene is met convergently, 130bp away, by another monocistron which encodes cell division protein FtsH (EC 3.4.24.-), which itself has a putative 6-gene operon transcribed divergently to it, 564bp away. This putative operon encodes the following proteins, in order of transcription:

1. exporter of the RND superfamily; mmpl family protein
2. -3bp (overlap) later, putative sigma E regulatory protein, MucB/RseB
3. 70bp later, 1 TMS, 474aa hypothetical protein with no significant homology
4. 152bp later, 69aa hypothetical protein: DUF2281 family protein
5. 0bp later, PIN domain protein
6. 93bp later, NifU-like domain protein

RegPredict analysis of the Spirochaetes was performed using the intergenic sequence preceding the putative Sec-independent protein translocase protein gene 27bp upstream of the FUPA32 gene in *T.denticola* (strains ATCC 35405 and F0402) and the equivalent sequence from *T.vincentii* ATCC 35580.

The first sequence predicted to regulate the FUPA32-encoding operon in *T.denticola* is found upstream of the translocase protein gene (as are all other predicted regulatory sequences for this operon), 73bp away. The sequence and score are: TtGaTACtg|C|ttGTAaCtA, score=7.55

There is only one member of this sequence set, although the set does have a duplicate with slightly different position frames, it is not described here in detail because the scores are slightly lower. The sequence (TtGctaCtg|C|ttGgcaCtA, score=5.96) is found 163bp upstream of a bicistron encoding ABC transporter, ATP-binding/permease protein, (TC 3.A.1.21.2) Fe^{3+}-carboxymycobactin transporter, IrtAB-like and ABC transporter, ATP-binding/permease protein, (TC 3.A.1.21.1) Fe^{3+}-Yersiniabactin uptake transporter-like.

The second sequence predicted to regulate the FUPA32-encoding operon is found 35bp upstream of it. The sequence and score are: TaCgccCTac|aaAGataGgA, score=7.65.

The first member of this motif set is a sequence (TaagAcCTgc|agAGaTagaA, score=5.06) found 143bp upstream of a bicistron encoding Fe-S oxidoreductases of moaA/nifB/pqqE family and a 848aa hypothetical protein 45% to cytochrome c biogenesis protein from *Helicobacter hepaticus* ATCC 51449 (length=936aa).

The second member of this motif set is a sequence (TaCcTcCtcc|aagGaAaGgA, score=5.06) found 135bp upstream of a monocistron encoding TPR domain protein,Tfp pilus assembly protein PilF.

The third member of this motif set is a sequence (TaCgGTCTtT|AcAGACaGgA, score=4.94) found 168bp upstream of a tricistron encoding the following proteins:

1. *grdE-2* encodes Glycine reductase component B beta/alpha subunits (EC 1.21.4.2) (Mg^{2+} required for decomposition of acetyl phosphate by protein C [Stadtman TC, 1989])
2. *grdB-2* encodes Glycine reductase component B gamma subunit (EC 1.21.4.2), selenocysteine-containing
3. Topoisomerase IV subunit A (EC 5.99.1.-)

The third sequence predicted to regulate the FUPA32-encoding operon is found 102bp upstream of it. The sequence and score are: AttGgagGAa|T|gTCaagCttT, score=7.25.

The first member of this motif set is a sequence (ttttTggaAa|T|gTcaaAcgtt, score=4.96) found 188bp upstream of a monocistron encoding Flagellar protein FlgJ [peptidoglycan hydrolase] (EC 3.2.1.-).

The second member of this motif set is a sequence (AttatAggAa|T|gTgaTtgttT, score=4.91) found 49bp upstream of an operon encoding 2 oligopeptide/dipeptide ABC transporter, permease proteins, 2 oligopeptide/dipeptide ABC transporter, ATP-binding proteins, 1 oligopeptide/dipeptide ABC transporter, periplasmic peptide-binding protein and 1 zinc carboxypeptidase family protein.

TC 3.A.3.32.6

There are two and four genes almost certainly coexpressed with those encoding **Msm4** and **Mst3**, respectively. Only one of these genes, encoding a Na^{+}-driven

multidrug efflux pump, is found near FUPA32 genes similar to these with any significant frequency. These are the only FUPA32-containing archaeans, and use of RegPredict is not available for archaeans yet.

Msm4 (*Methanobrevibacter smithii* 707aa, gi:222444459) is encoded in a bicistron, following a gene encoding an 86aa hypothetical protein with no significant homology 11bp upstream. Eighteen base pairs downstream of the FUPA32 gene, convergently transcribed, is a single gene encoding CAAX amino terminal protease family protein. This gene in turn has a bicistron transcribed divergently from it, 170bp away, the first encoding ABC-type nitrate/sulfonate/bicarbonate transport system, permease component and the second, 11bp away, encoding the ATPase component of the same transport system.

Divergently transcribed from the FUPA32-encoding operon, a single gene 175bp away is found which encodes formate dehydrogenase chain D (EC 1.2.1.2) (tungsten and iron are both found in this enzyme in *Methylobacterium extorquens* [Laukel M, et al., 2003] but “no divalent metal cation is needed as a cofactor for enzyme activity” in *Gelatoporia subvermispora*, a eukaryote [Watanabe T, et al., 2008]). A tricistron is transcribed convergently to this gene, 3bp away. This operon encodes LSU ribosomal protein L15e, followed 31bp later by exosome subunit, DUF54 family protein and then 60bp later, 4-carboxymuconolactone decarboxylase (CMD) family protein (EC 4.1.1.44) (no known metal requirements). Divergently transcribed from this tricistron, 198bp away, a 6-gene putative operon is found which encodes first a DUF3795 protein likely to

be zinc binding given its conserved cysteines and 42% similarity to the C-terminal half of GCN5-related N-acetyltransferase (GNAT) family proteins, followed 271bp later by DppABCDF, the substrate binding protein, permeases and ATP-binding components of a dipeptide ABC transport system (TC# 3.A.1.5.2).

Mst3 (*Methanosphaera stadtmanae* 705aa, gi: 84489626) is encoded by the fourth gene in a putative 7 gene operon, although the intergenic space between the latter genes reduces the likelihood that they are cotranscribed. This putative operon encodes the following proteins, in order of transcription:

1. transcriptional regulator MarR family
2. -28bp (overlap) later, Multi antimicrobial extrusion protein (Na^+:drug antiporter), MATE family of MDR efflux pumps
3. 511bp later, 1 TMS, 103aa hypothetical protein with no significant homology
4. 65bp later, FUPA32

*the remaining proteins in this list are not predicted to be coexpressed but have been included for completeness

5. 677bp later, 159aa hypothetical protein with no significant homology
6. 474p later, 3 TMS, 141aa hypothetical protein with no significant homology
7. 121bp later, Potassium uptake protein TrkH

Transcribed convergently to the TrkH-encoding gene, 69bp away, there is a 4-gene operon, which encodes two copies of potassium voltage-gated channel subfamily KQT, 83bp apart, followed, with a 37bp overlap, by 180aa hypothetical protein with no

significant homology and then 140aa CRISPR-associated Csh1 family protein, which also overlaps its preceding gene, by 16bp. Two genes are transcribed divergently from this operon, 190bp away, which may originate from a single, very large gene encoding a cell surface protein. The first gene encodes an Asn/Thr-rich adhesin-like protein domain and the second, 427bp away, encodes Asn/Thr-rich surface protein. The shared PSIBLAST hits as high asparagine/threonine proteins suggest these proteins are closely related.

Divergently transcribed from the FUPA32-encoding operon, 124bp away from the MarR-encoding gene, a single gene encoding a 148aa hypothetical protein with no significant homology is found.

The archaeans encoding FUPA32 proteins are not available for analysis in RegPredict.

TC 3.A.3.32.8

Cyanobacteria

Tel5 (*Thermosynechococcus elongatus* BP-1 769aa, gi:22295937) is encoded 24bp downstream, codirectionally, from a bestrophin superfamily protein, implicated in Cl^- transport (Chloride channels: This superfamily of poorly-understood channels consists of approximately 13 members. They include ClCs, CLICs, Bestrophins and CFTRs. “These channels are non-selective for small anions; however chloride is the most abundant anion, and hence they are known as chloride channels” [SinghH, 2010; Hagen AR, et al., 2005]). Two other proteins are encoded downstream of Tel5 codirectionally. First, 56bp

downstream, a phage integrase is encoded, followed 160bp later by a XisA-like site specific C-terminal recombinase from phage integrase family. This gene is met convergently with a 22bp overlap by the last gene in a 3-gene operon. The operon encodes chaperone protein DnaK, followed 13bp later by an 87aa hypothetical protein in the DUF3146 superfamily and lastly, 66bp later, a 1 TMS, 67aa hypothetical protein with no significant homology. Upstream of the bestrophin gene one gene is transcribed divergently, 187bp away, encoding CAB/ELIP/HLIP superfamily protein, putatively involved in light-harvesting. Two genes are transcribed convergently to this gene, the end of the second one 633bp away. These genes encode UDP-glucose 4-epimerase (EC 5.1.3.2) (contains none of the common metal ions [Arabshahi A, et al., 1988]) and 37bp later a 1 TMS, 93aa hypothetical protein with no significant homology.

Tel2 (*T.elongatus* BP-1 769aa, gi:22294378) is encoded by a gene with no other co-directional genes adjacent to it. There are two genes transcribed divergently from it, starting 228bp away. These genes encode enolase (EC 4.2.1.11) (two Mg^{2+} per subunit are required for catalytic activity [Hosaka T, et al., 2003]) followed 1bp later by SSU rRNA (adenine(1518)-N(6)/ adenine(1519)-N(6))-dimethyltransferase (EC 2.1.1.182) (Mg^{2+} required, NaCl and KCl slightly stimulate up to 0.2 M [Andrésson OS, et al., 1980]). A 5- or 6-gene operon is transcribed convergently with the FUPA32 gene, its last gene overlapping the FUPA32 gene by 13bp. The putative operon encodes the following proteins in order of transcription:

1. resolvase RNase H domain fold-containing protein

2. 39bp later, exoribonuclease II (EC 3.1.13.1) (Mg^{2+} and K^{+} are required for activity [Bollenbach TJ, et al., 2004])
3. 11bp later, hypothetical protein in the FliB superfamily, 74% homologous to Fe-S cluster protein (gi:359464215); referred to as ORF_ID:tlr0653
4. 46bp later, 117aa hypothetical protein referred to as ORF_ID:tlr0654
5. 28bp later, ABC transporter ATP-binding protein, (TC 3.A.1.106.-), probably xenobiotic transporter, e.g., multidrug efflux pump
6. 369bp later, ribosomal small subunit pseudouridine synthase A (EC 4.2.1.70) (dependent on presence of Mg^{2+}, Co^{2+}, Fe^{2+} or Mn^{2+}, inhibited by Zn^{2+} and Ni^{2+} [Heinrikson RL, et al., 1964; Preumont A, et al., 2008]); RsuA-16S rRNA pseudouridylate 516 synthase

The 2-gene operon divergently transcribed from the FUPA32 gene is met convergently by 6 or 7-gene operon whose last gene overlaps it by 14bp. The putative operon encodes the following proteins in order of transcription:

1. dihydroorotase (EC 3.5.2.3) (divalent cation required, predicted zinc-metalloprotein because activity varies by cation in the order: $Zn^{2+} > Co^{2+} > Cd^{2+}$ [Porter TN, et al., 2004])
2. 526bp later, dTDP-glucose 4,6-dehydratase (EC 4.2.1.46) (Mg^{2+} may be required [Singh B, et al., 2010])
3. 15bp later, UDP-glucose dehydrogenase (EC 1.1.1.22) (no metal ion requirement data for prokaryotes)

4. 3bp later, ribonuclease HII (EC 3.1.26.4) (requires Mg^{2+} or Mn^{2+} to function, and in presence of 10 mM of Ba^{2+}, Ca^{2+}, Co^{2+}, Zn^{2+}, Cu^{2+}, Fe^{2+}, or Sr^{2+} [Ohtani N, et al., 2000])
5. -19bp (overlap) later, phenylalanyl-tRNA synthetase beta chain (EC 6.1.1.20) (requires Mg^{2+} [Barrett AR, et al., 2008])
6. -54bp (overlap) later, biotin synthase (EC 2.8.1.6) (the enzyme contains an iron-sulfur cluster and Fe^{2+} enhances activity, as does S^{2-} [Kiyasu T, et al., 2002])
7. 186bp later, putative GIDE superfamily E3 ubiquitin ligase

Most or the convergently transcribed genes flanking the transcription unit containing the FUPA32 gene occur near the gene at too low of a frequency to suggest that their expression is joined, but the two nearest the FUPA32 gene occur at about the same frequency as the two divergently transcribed from it.

Ava3 (*A.variabilis* ATCC 29413 737aa, gi:75907215) and **Nsp6** (*Nostoc* sp. PCC 7120 735aa, gi:17230400) are encoded in very similar operons and as such will be described together.

Ava3 is encoded as the twelfth gene of 14 transcribed codirectionally and thus may be part of a very large operon. The putative operon encodes the following proteins, in order of transcription:

1. soluble [2Fe-2S] ferredoxin
2. -10bp (overlap) later, Staphylococcus nuclease (SNase) domain

3. 80bp later, Inositol-1-monophosphatase (EC 3.1.3.25) (Mg^{2+} required, optimal activation effect at 3 mM [Patra B, et al., 2007])
4. 257bp later, COG2214: DnaJ-class molecular chaperone
5. 185bp later, ATP phosphoribosyltransferase regulatory subunit (EC 2.4.2.17) (requires Mg^{2+} [Lohkamp B, et al., 2004])
6. 182bp later, 4Fe-4S ferredoxin, iron-sulfur binding
7. 242bp later, 1 TMS, 261aa hypothetical protein 61% homologous to a 200aa stretch of the last third of a 3kDa Na^{+}-Ca^{2+} exchanger/integrin-beta4 (gi:186465199)
8. 81bp later, branched-chain amino acid transport ATP-binding protein LivF (TC 3.A.1.4.1)
9. 77bp later, glutamate receptor 1 precursor
10. 641bp later, 2 copies of carbonic anhydrase (EC 4.2.1.1) (zinc metalloenzyme, Co^{2+} can substitute for Zn^{2+} [Elleby B, et al., 2001]), 375bp apart.

*The second of these more closely resembles putative phosphoribosyl-AMP

11. cyclohydrolase in *N.sp*
12. 396bp later, FUPA32
13. 218bp later, Mov34/MPN/PAD-1 family
14. 64bp later, sulfur carrier protein adenylyltransferase ThiF

A 2-gene putative operon convergently meets this operon as described in *A.vaiabilis*, 318bp away. It encodes FOG: HEAT repeat followed 19bp later by

transposase. A homologue of the transposase is also encoded divergently from the FOG: HEAT repeat gene.

In *N.sp.* however, the FUPA32-encoding operon continues for two more genes, 137bp downstream, encoding a 189aa hypothetical protein with no significant homology followed 62bp later by permease of the major facilitator superfamily. These same genes are found on the other side of the above mentioned transposase sandwich, which was presumably inserted here in *A.variabilis* after these species split.

A 6-gene putative operon is also transcribed divergently from that encoding the FUPA32 protein, 670bp away. It encodes, in order of transcription:

1. 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) (binds NADP+, does not use metal cofactor [Silva RG; et al.,2006, 2008])
2. 489bp later, tRNA-(ms[2]io[6]A)-hydroxylase (EC 1.-.-.-)
3. 18bp later, lactoylglutathione lyase-type protein
4. 250bp later, histone acetyltransferase HPA2-type protein
5. 7bp later, competence protein F homolog, phosphoribosyltransferase domain; protein YhgH required for utilization of DNA as sole source of carbon and energy
6. 69bp later, RTX toxins and related Ca^{2+}-binding proteins

In *N.sp*, this operon is interrupted by two gene transcribed in the opposite direction located between 3 & 4, encoding HigBA toxin/antitoxin addictive gene pair. The operon is in turn met convergently by a 4-gene operon, 207bp away which encodes the following protein in order of transcription:

1. glyoxalase family protein
2. 524bp later, cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases
3. 29bp later, carbonic anhydrase (EC 4.2.1.1) (zinc metalloenzyme, Co^{2+} can substitute for Zn^{2+} [Elleby B, et al., 2001])
4. 174bp later, RNA-binding protein

   This is the same in both species.

Ava18 (*A.variabilis* ATCC 29413 770aa, gi:75911260) and **Nsp5** (*Nostoc* sp. PCC 7120 771aa, gi:17229496) are also encoded in very similar operons and as such will be described together. **Ava18** is encoded as the fifth gene of 7 transcribed codirectionally as one putative operon. The putative operon encodes the following proteins, in order of transcription:

1. serine protease, DegP/HtrA, do-like (EC 3.4.21.-)
2. 314bp later, peptide deformylase (EC 3.5.1.88) (native protein contains Fe^{2+}, Co^{2+} can fully substitute. Ni^{2+} can substitute as well up to 0.1 mM, higher concentrations become inhibitory. Zn^{2+} can substitute but only poorly, not sensitive to oxidation like the native enzyme is [Yen NT, et al., 2010; Rajagopalan PTR, et al., 1998])
3. 92bp later, 386aa (383aa in *N.sp.*) FUPA32 P-type ATPase fragment
4. 36bp later, 236aa (151aa in *N.sp.*) FUPA32 P-type ATPase fragment
5. 104bp later, FUPA32

6. 75bp later, 1 TMS, 100aa hypothetical protein
7. 823bp later, 195aa hypothetical protein with 43% similarity to FUPA32 N-terminus over 77aa

This operon is met convergently, 98bp away, by a single gene encoding omega amidase (Nit2 homologue) which itself has 1 gene transcribed divergently to it, 248bp away, encoding multiple antibiotic resistance protein marC.

In *A.variabilis*, no genes are found upstream of the FUPA32-encoding operon for at least 17.5kb and as such none with be discussed. In *N.sp.*, two genes are transcribed divergently to the aforementioned operon, starting 302bp away. The first gene in this bicistron encodes chromosomal replication initiator protein DnaA followed 415bp later by DNA polymerase II beta subunit (EC 2.7.7.7) (the holoenzyme requires Mg^{2+} and also Zn^{2+} [Kornberg T, et al., 1974 & Setlow P, 1974]). This operon is met convergently, 103bp away, by a 3-gene operon encoding type 12 methylase, followed 6bp later by oligopeptide transport system permease protein OppB and OppC (TC 3.A.1.5.1), 4bp apart. This short operon in turn has another 3-gene operon transcribed divergently from it, 361bp away. This operon encodes 150aa hypothetical protein in DUF3531 superfamily, followed 8bp later by MutT/nudix family protein and lastly, 211bp later, a 97aa hypothetical protein with no significant homology.

RegPredict Analysis of genes putatively coregulated with FUPA32 in Cyanobacteria was performed using intergenic sequence preceding the Bestrophin (putative Cl^- channel) superfamily gene for Tel5, sequence directly upstream for Tel2, sequence directly upstream of Ava3 and Nsp6, and sequence both preceding serine

protease, DegP/HtrA, do-like (EC 3.4.21.-) gene and sequence directly upstream for Ava18 and Nsp5.

The first predicted regulatory motif is found preceding all four FUPA32 genes (treating Ava3/Nsp6 and Ava18/Nsp5 as one gene each). This motif can be summarized, with sequences scoring according to their similarity to it, as (tCTncnntnGc|nCnnngtcAGc). Capital letters indicate consistent palindromic sequence, lower case letters indicate consistent nucleotides without reliable palindromic compliments. The sequences are as described below, arranged by their associated FUPA32 protein and including the nearest downstream gene and distance from it.

Tel5	bestrophin gene	59bp	AgcGcCGTttc|aacACGcCagT, score=4.53
Tel2	*tel2*	54bp	aCTAcCTAgAc|cTtTAGtTAGc, score=4.35
Ava3/Nsp6	*ava3*	110bp	aCtAtggTAGg|tCTAgtcTtGc, score=3.99
	nsp6	116bp	TCctcaCAAGc|aCTTGattaGT, score=4.47
Ava18/Nsp5	*ava18*	96bp	TCTGcaTtctc|tcagAgtCAGA, score=5.02
	nsp5	99bp	TCTGcatTctc|tcaAggtCAGA, score=5.10

The first member of this motif set is a sequence (TaTAAAatctc|actgaTTTAgA, score=4.20) found only in *Nostoc sp*., 174bp upstream of an operon encoding the following proteins:

1. serine/threonine kinase
2. tRNA-i(6)A37 methylthiotransferase

3. *leuB* encodes 3-isopropylmalate dehydrogenase (EC 1.1.1.85) (divalent cation required, at 0.5 mM Mn^{2+} causes optimal stimulation followed by Mg^{2+} and Co^{2+}; Ca^{2+} and Ba^{2+} inhibit [Wallon G, et al., 1997])
4. 88aa hypothetical protein with no significant homology
5. Leader peptidase (Prepilin peptidase) (EC 3.4.23.43) (no known metal requirements)/ N-methyltransferase (EC 2.1.1.-)

The second member of this motif set is a sequence (GCTAtAgTAgc|ttTAgTtTAGC, score=4.63) found only in *T.elongatus*, 212bp upstream of an operon encoding the following proteins:

1. Ribosomal large subunit pseudouridine synthase E (EC 4.2.1.70) (dependent on presence of Mg^{2+}, Co^{2+}, Fe^{2+} or Mn^{2+}, inhibited by Zn^{2+} and Ni^{2+} [Heinrikson RL, et al., 1964; Preumont A, et al., 2008])
2. heterocyst specific ABC-transporter, membrane fusion protein DevB homolog
3. devC-like ABC transporter permease protein
4. ATP-binding protein of devA-like ABC transporter

The third member of this motif set is a sequence (tcTatagTAGc|tCTAtatcAat, score=4.25) found only in *A.variabilis*, 27bp upstream of an bicistron encoding Inactive homolog of metal-dependent proteases, putative molecular chaperone and 5'-nucleotidase surE (EC 3.1.3.5) (Mg^{2+}, Mn^{2+} and Co^{2+} can all activate the enzyme. Ni^{2+} may as well [Proudfoot M, et al., 2004] or it may inhibit it [Itami H, et al., 1989]).

The fourth member of this motif set is a pair of sequences (tgcacCtTttT|AttAcGtcttt, score=4.30) found both in *A.variabilis* and *Nostoc sp.*, 50bp

upstream in both genomes, of a monocistron encoding Transcriptional Regulator, Crp/Fnr family.

The fifth member of this motif set is a sequence (TgcGcCtCTac|tcAGgGaCagA, score=4.48) found only in *T.elongatus*, 108bp upstream of an operon encoding the following proteins:

1. Secretion protein HlyD
2. Glyoxalase/bleomycin resistance protein/dioxygenase
3. Cytosol aminopeptidase PepA (EC 3.4.11.1) (Mg^{2+} activates [Mathew Z, et al., 2000])
4. Ribosomal RNA small subunit methyltransferase E (EC 2.1.1.-)
5. *rps12* encodes SSU ribosomal protein S12p (S23e)
6. *rps7* encodes SSU ribosomal protein S7p (S5e)
7. *fus* encodes Translation elongation factor G
8. *tufA* encodes Translation elongation factor Tu
9. *rps10* encodes SSU ribosomal protein S10p (S20e)
10. Peptidase M61
11. *uvrC* encodes excinuclease ABC subunit C

The sixth member of this motif set is a sequence (TacacCCttac|acccGGtctcA, score=4.30) found only in *A.variabilis*, 164bp upstream of a tricistron encoding:

1. putative Crp/Fnr family transcriptional regulator
2. Carbonic anhydrase (EC 4.2.1.1) (zinc metalloenzyme, Co^{2+} can substitute for Zn^{2+} [Elleby B, et al., 2001])

3. *rbpB* encodes RNP-1 like RNA-binding protein

The seventh and last member of this motif set is a sequence (TacacCcttac|acccaGtctcA, score=4.30) found only in *Nostoc sp.*, 102bp upstream of a monocistron, moaE, which encodes molybdenum cofactor biosynthesis protein MoaE.

The second predicted regulatory motif is found preceding three FUPA32 genes (treating Ava3/Nsp6 and Ava18/Nsp5 as one gene each). This motif can be summarized as (GaTGTnnAt|ntnnnnnnC). The sequences are as described below, arranged by their associated FUPA32 protein and including the nearest downstream gene and distance from it.

Tel5	bestrophin gene	105bp	GatttTCTT|AAGAtttgC, score=4.07
Ava3/Nsp6	*ava3*	67bp	GaTAtTGAa|gTCAtTAcC, score=4.32
Ava18/Nsp5	*ava18*	69bp	GacgtgCAt|tTGtcctcC, score=4.46
	nsp5	72bp	GatGtgCAt|cTGtcCtcC, score=4.79

The first member of this motif set is a pair of sequences (GttGtaGAT|ATCatCgcC, score=4.33; GatgtaGAT|ATCatcgcC, score=4.48) found both in *A.variabilis* and *Nostoc sp.*, 61bp upstream in both genomes, of a monocistron, *pbpB*, which encodes multimodular transpeptidase-transglycosylase (EC 2.4.1.129) (slight stimulation with Mg^{2+}, but no divalent cation required for function [Nakagawa J, et al., 1984]), (EC 3.4.-.-).

The second member of this motif set is also a pair of sequences (GtTttgtAa|aTgtttAcC, score=4.12 in both) found both in *A.variabilis* and *Nostoc sp.*,

221 and 223bp upstream in both genomes, of a monocistron, *argJ*, encoding glutamate N-acetyltransferase (EC 2.3.1.35) (no known ion requirements)/ N-acetylglutamate synthase (EC 2.3.1.1) (no known ion requirements)

The third member of this motif set is a sequence (GaTgcTgtt|gtgActAcC, score=4.05) found only in *A.variabilis*, 190bp upstream of a bicistron which encodes two copies of a ABC transporter of aliphatic sulfonates, substrate-binding component.

The fourth and last member of this motif set is a sequence (GatgAtCAt|tTGcTgtcC, score=4.05) found only in *Nostoc sp.*, 82bp upstream of a 14-gene operon, which encodes the following proteins:

1. Protein of unknown function DUF433
2. Protein of unknown function DUF82
3. CRISPR-associated protein Csc3
4. CRISPR-associated protein Csc2
5. CRISPR-associated protein Csc1
6. CRISPR-associated helicase Cas3
7. 205aa hypothetical protein with 44% similarity to Fe^{2+}-dependent oxygenase superfamily
8. CRISPR-associated protein Cas6
9. CRISPR-associated RecB family exonuclease Cas4b
10. CRISPR-associated protein Cas1
11. Transposase
12. CRISPR-associated protein Cas2

13. 101aa hypothetical protein with no significant homology
14. unknown CRISPR-associated protein

The third predicted regulatory motif is found preceding two FUPA32 genes (treating Ava18/Nsp5 as one gene). This motif can be summarized as (nGAnnTnCnT|AnGntnTCCa). The sequences are as described below, arranged by their associated FUPA32 protein and including the nearest downstream gene and distance from it.

Tel5	bestrophin gene	106bp	gGatttTCTT|AAGAtttgCa, score=4.85
Ava18/Nsp5	*ava18*	70bp	AGacGtgCAt|tTGtcCtcCT, score=4.69
	nsp5	73bp	AGatGtgCAt|cTGtcCtcCT, score=4.77

The first member of this motif set is another close sequence (aGatgAtCAt|tTGcTgtcCa, score=4.20) found 83bp upstream of the CRISPR-associated protein operon in *N.sp* and thus will not be repeated.

The second member of this motif set is a sequence (gGatttGcaT|AatCtctaCa, score=4.65) found only in *A.variabilis*, 62bp upstream of an operon encoding the following proteins:

1. Ferrichrome-iron receptor
2. Dipeptide-binding ABC transporter, periplasmic substrate-binding component (TC 3.A.1.5.2)
3. Oligopeptide transport system permease protein oppB
4. Oligopeptide transport system permease protein oppC
5. Oligopeptide ABC transporter, ATP-binding protein

6. Oligopeptide transport ATP-binding protein oppF
7. Ribonuclease H-like
8. Isocitrate dehydrogenase

The third member of this motif set is a sequence (gtAttttCaT|AgGttctTca, score=4.33) found only in *Nostoc sp.*, 225bp upstream of a bicistron encoding phycocyanobilin:ferredoxin oxidoreductase PcyA (EC 1.3.7.5) (no known metal requirements) and a 1 TMS, 62aa hypothetical protein with no significant homology.

The fourth member of this motif set is a pair of sequences (gGaAttTCta|ctGAttTaCa, score=4.02; gGaAttTgta|ctgAttTaCa, score=4.48) found both in *A.variabilis* and *Nostoc sp.*, 95bp upstream in both genomes, of a bicistron encoding Type II secretory pathway, ATPase PulE/Tfp pilus assembly pathway, ATPase PilB and WD-40 repeat protein.

The fifth member of this motif set is a sequence (AGATtttccT|AttcttATCT, score=4.22) found only in *A.variabilis*, 39bp upstream of a tricistron encoding Glucose-methanol-choline (GMC) oxidoreductase:NAD binding site and Cytochrome d ubiquinol oxidase subunits I and II (EC 1.10.3.-).

The fourth predicted regulatory motif is found preceding two FUPA32 genes (treating Ava3/Nsp6 as one gene). This motif can be summarized as (nCAAagTCn|nGAnTTnna). The sequences are as described below, arranged by their associated FUPA32 protein and including the nearest downstream gene and distance from it.

Tel2	*tel2*	72bp	acaaAGTCg|gGACTaata, score=4.91
Ava3/Nsp6	*ava3*	126bp	GcaAAcaCt|tGatTTaaC, score=4.21
	nsp6	11bp	acaagcaCt|tGattagtc, score=3.89

The first member of this motif set is a sequence (gcaaActCT|AGgcTaata, score=4.11) found only in *T.elongatus*, 142bp upstream from a 4-gene operon encoding the following proteins:

1. *rps1* encodes SSU ribosomal protein S1p
2. *pdxA* encodes 4-hydroxythreonine-4-phosphate dehydrogenase (EC 1.1.1.262) (Zn^{2+} tightly bound, divalent metal ion-dependent enzyme, can be replaced by other divalent cations, e.g., Mg^{2+} [Banks J, et al., 2004])
3. 107aa hypothetical protein in DUF1825 superfamily (Cyanobacteria-specific)
4. Branched-chain amino acid aminotransferase (EC 2.6.1.42) (no known metal ion required) / Aminodeoxychorismate lyase (EC 4.1.3.38) (no known metal ion required), IlvE/PabAc.

The second member of this motif set is a pair of sequences (TcaAacTct|tcAccTatA, score=4.17 in both) found both in *A.variabilis* and *Nostoc sp.*, 84bp upstream in both genomes, of an operon encoding the following proteins:

1. *hoxU* encodes NAD-reducing hydrogenase subunit HoxU (EC 1.12.1.2) ([NiFe] hydrogenase, contains nickel and iron in the catalytic core [Rangarajan ES, et al., 2008])
2. Unidentified ORF in hydrogenase gene cluster
3. *hoxY* encodes NAD-reducing hydrogenase subunit HoxY (EC 1.12.1.2)

4. Inosine-5'-monophosphate dehydrogenase related protein
5. NAD-reducing hydrogenase subunit HoxH (EC 1.12.1.2) (*A.variabilis* only)

The third member of this motif set is also a pair of sequences (ataAagTtt|ggAtaTata, score=4.12; ataAagTTt|gAAtaTata, score=3.82) found both in *A.variabilis* and *Nostoc sp.*, 175 and 170bp upstream, respectively, of an operon encoding the following proteins:

1. transposase (*Nostoc sp.* only)
2. Phosphoadenylyl-sulfate reductase [thioredoxin] (EC 1.8.4.8) (iron-sulfur protein [Bhave DP, et al., 2011])
3. Exonuclease SbcD
4. Protein pucC
5. Histidinol-phosphatase [alternative form] (EC 3.1.3.15) (Fe^{2+}, Zn^{2+} part of the trinuclear metal center [Omi R, et al., 2007])

The fourth member of this motif set is also a pair of sequences (agaAacTct|caAccTaaa, score=3.76; acaAacTct|caAccTaaa, score=4.46) found both in *A.variabilis* and *Nostoc sp.*, 55bp upstream in both genomes, of an operon encoding the following proteins:

1. *nifV1* encodes Homocitrate synthase (EC 2.3.3.14) (no known metal requirements in prokaryotes)
2. *nifZ* encodes NifZ nitrogenase MoFe maturation protein

3. *nifT* encodes NifT nitrogenase MoFe maturation protein ("induced not only in a nitrogen-depleted culture but also by iron depletion irrespective of the nitrogen status" [Stricker O, et al., 1997])
4. NADH:ubiquinone oxidoreductase, NADH-binding (51 kD) subunit

The fifth and final member of this motif set and the Cyanobacterial group, is a sequence (TAaAAgTCt|cGAtTTaTA, score=4.14) found only in *A.variabilis*, 175bp upstream of a monocistron encoding Nitrogenase (vanadium-iron) alpha/delta chains (EC 1.18.6.1) (Mo and Fe required, 18-36 atoms of iron per molecule of MoFe protein, Mg^{2+} required for MgATP complex [Eady RR, et al., 1974]).

Upon analysis of recurring genes predicted to be expressed with FUPA32 genes, several genes stand out as prime candidates for functional prediction. More than any other gene, heavy-metal associating proteins highly similar to the N-terminal HMA domains of FUPA32 proteins were found in 1 to 3 copies in every species analyzed except the Archaeans and *T.elongatus*. These are found both upstream and downstream of the complete FUPA32 genes, implying that their abundance is significant. They may act as chaperones of the substrate cation to the FUPA32 transporters. Putative oxidoreductases are encoded adjacent to Rpa7, Aar5 and Wsu4, as are Fe-binding ferritin-like proteins, multiple ferredoxins and Fe-responsive regulators, eg, Irr. Additionally, ferroxidase, superoxide dismutase[Fe], high affinity iron permease and periplasmic p19 protein involved in high affinity iron transport are also found co-localized near Wsu4. Cytochrome c4 and c5 and ferredoxin proteins IscUA are found co-

localized and coregulated, respectively, with Aar5. Cytochrome c4 and c5 are also coregulated, in 3 distinct instances, with Rpa7 as is cytochrome P450 hydroxylase. The iron-requiring (or preferring) proteins dihydroxy-acid dehydratase (EC 4.2.1.9), geranyltranstransferase (farnesyldiphosphate synthase) (EC 2.5.1.10), formate dehydrogenase chain D (EC 1.2.1.2), methionine sulfoxide reductase MsrB (EC 1.8.4.12) and enoyl-CoA hydratase (EC 4.2.1.17) are also found to be coexpressed with FUPA32 in Proteobacteria.

In the Spirochaete *T.denticola*, a FUPA32 N-terminal HMA domain fragment is found associated with the complete FUPA32 gene, as is the ferredoxin assembly protein gene NifU. Moreover, 2 ABC transporter ATP-binding/permease proteins (TC 3.A.1.21.2), which are Fe^{3+}-siderophore importers, as well as Fe-S oxidoreductase and a putative cytochrome c biogenesis protein are found to be coregulated with this FUPA32.

In the archaeans, no coregulatory analysis was possible, but of the two most reliably co-localized cistrons associated with these FUPA32 genes, one is a formate dehydrogenase chain D protein (EC 1.2.1.2), found to contain Fe and sometimes tungsten (Laukel M, et al., 2003).

In Cyanobacteria, two soluble ferredoxin proteins are co-localized with Ava3/Nsp6 and one, FliB is co-localized with Tel2. The Fe^{2+}-containing peptide deformylase (EC 3.5.1.88) is also co-localized with Ava18/Nsp5. Additionally, Ribosomal large subunit pseudouridine synthase E (EC 4.2.1.70), which prefers Fe^{2+} among its cofactors, is found to be coregulated with Tel2. In *Nostoc sp.*, Fe^{2+}-dependent oxygenase superfamily protein, phycocyanobilin:ferredoxin oxidoreductase PcyA (EC

1.3.7.5) and Phosphoadenylyl-sulfate reductase [thioredoxin] (EC 1.8.4.8) are coregulated with the FUPA32 genes. In *A.variabilis*, ferrichrome-iron receptor and Nitrogenase (vanadium-iron) alpha/delta chains (EC 1.18.6.1) are predicted to be coregulated with the FUPA32 genes. Lastly, In both *Nostoc sp.* and *A.variabilis*, the [NiFe] core-containing NAD-reducing hydrgenase (EC 1.12.1.2) subunits HoxUY and the nitrogenase [MoFe] core-containing protein (EC 2.3.3.14) subunits NifVZT, which are induced by iron depletion are coregulated with FUPA32 genes. With this wealth of Fe-responsive, Fe-importing and Fe-containing proteins coexpressed with the FUPA32 genes, we confidently suggest that FUPA32s function as Fe^{2+} importers, very possibly in a direct-delivery capacity to ferredoxin assembly proteins.

Discussion

The prokaryotic FUPAs appear to function as condition-specialized, stress environment-type transporters, largely predicted to participate in necessary cation influx. Functional prediction and putative characterization has shed light on likely functions of many of these families, and provides a path for future research.

FUPA23 is predicted to act as a Mg^{2+} uptake transporter, providing the magnesium cation to proteins that require it for activity. The genes encoding these ATPases are predicted to be involved in cell division or antibiotic resistance. This prediction is supported by the wealth of metalloproteins found to be coexpressed with FUPA23s. Many prefer or require Mg^{2+} for activity, and of those proteins, about one quarter are involved in cell division or antibiotic resistance. This prediction is in keeping with the type II topology of the family, and would represent a second family of Mg^{2+} P-type ATPase transporters.

FUPA24 ATPases are also predicted to act as Mg^{2+}/divalent cation uptake transporters, specifically expressed within host cells in most cases. These proteins are predicted to be involved in the biosynthesis of CoenzymeA, cytochrome c and lysyl-/isoleucyl-tRNAs. This prediction is derived from the observation that over three fourths of the metalloproteins co-expressed with FUPA24s prefer or require Mg^{2+} and that about a quarter of these are involved in the aforementioned processes. The smaller family size and lower overall frequency of coexpressed metalloproteins, combined with the fact that this family is of the type I topology and thus typically known to be restricted to transport

of heavy metals, makes this prediction somewhat less likely. It is likely that the function of this family will become forthcoming as it is expanded. The function of the N-terminal half of the protein is not known. It has lost P-type ATPase activity, based on sequence analyses, showing that essential conserved residues are not present.

FUPA25 is predicted to act as a divalent cation uptake transporter, but there is little support for even this tentative prediction. These genes are predicted to be involved in cell division regulation or cell division-associated protein expression due to the fact that that one-quarter of the proteins co-expressed with FUPA25 are involved with cell division. Because the proteins co-expressed with members of this family do not appear to have cation preferences, no specific cation predictions can be made.

FUPA26 and FUPA32 are predicted to act as $Fe^{2/3+}$/siderophore uptake transporters. These systems could provide the iron cation to proteins that require it for activity. FUPA26 proteins appear to be involved in heme and cytochrome biogenesis in Actinobacteria due to their predicted co-expression with known proteins involved in these processes. Over half of the transporter proteins co-expressed with FUPA26 are iron or iron-complexed siderophore transport proteins, and about one quarter of the proteins co-expressed with FUPA26 in Actinobacteria are heme, siroheme, or cytochrome-related proteins. FUPA32 proteins appear to be involved in ferredoxin and ferrichrome iron provision in Spirochaetes, Proteobacteria, Flavobacteria and Cyanobacteria. This prediction is supported by the fact that over one third of the transport proteins co-expressed with FUPA32 homologues are iron or iron-complexed siderophore import proteins, and about one half of the proteins co-expressed with FUPA32 are

ferredoxin/Fe-S cluster proteins or proteins related to the assembly of ferredoxin. These predictions can be readily tested using genetic manipulation experiments, deleting either these P-type ATPases, the coexpressed putative iron transporters or both and exposing the mutants to varying iron concentrations. Coexpression may serve to provide both low affinity/high efficiency and high affinity/low efficiency transporters of the iron cation.

FUPA27 and 29 are predicted to act as Cu^{2+} insert transporters, providing the copper cation to the cytochrome c oxidase complex (Cco) or Nitrogen fixation (Fix) proteins embedded in the membrane. This prediction is strongly supported by the fact that over 95% of these proteins are found within highly conserved *cco/fix* operons and that most of the co-regulated proteins are either cytochrome or nitrogen fixation-related.

Hassani, et al. (2010) provided strong experimental evidence for this prediction. In this paper, they examined a putative copper translocating P-type ATPase, CtpA (TC# 3.A.3.27.4) which appears incapable of providing copper tolerance, and therefore seems to lack the capacity to efficiently transport copper across the membrane from the cytoplasm to the extracytoplasmic space. This was also shown to be the case because they demonstrated that mutants with CtpA deleted were actually rescued by increased medium copper concentrations, rather than decreased concentrations as would be expected if the transporter were involved in copper tolerance.

Two earlier reports (Kahn et al., 1989; Preisig et al., 1996) had suggested that similar ATPases, FixI in *Rhizobium meliloti* and *Bradyrhizobium japonicum,* serve a comparable role in the nitrogen fixation complex, FixGHIS, which is involved in the biogenesis of heme-copper oxidases including those involved in symbiotic nitrogen

fixation. These homologous complexes are believed to be involved in the assembly of membrane (cytochrome oxidase, cbb3), periplasmic (nitrous oxide reductase, NosZ) copper enzymes. The FUPA27/29 family ATPases, referred to alternatively as CtpA or FixI, appear to deliver copper to the active sites of these enzymes without effectively translocating the copper across the membrane. On the basis of our genome context analyses as well as the work reported by Hassani et al. (2010), we conclude that copper ATPases within the FUPA27 and FUPA29 families function to deliver copper to enzymes dependent on copper for activity. Further experiments in this area involving cytochrome c and nitrogen fixation functions in the absence of these proteins would make a significant contribution to the characterization of these proteins.

FUPA28s are predicted to act as Zn^{2+} and Fe^{2+} uptake transporters, harvesting the cations from within their hosts. The only organisms known to possess them are intracellular parasites of the *Legionella* genus. These transporters are consistently found expressed with intracellular multiplication factor ICM-type IV protein secretion genes, as well as being found to be coexpressed with FUR/ZUR regulatory proteins, Fe^{2+} transport proteins and hemeoxygenases. This may indicate that these genes function strictly upon entry into the host cell, an unusual trait for P-type ATPases, which are usually thought to maintain homeostasis and known to be some of the first genes to be eliminated in the genome reduction process associated with parasitism (Chan, et al., 2010).

FUPA30s are predicted to act as Ca^{2+} efflux transporters, removing the calcium cation from the cytoplasm at appropriate times to deactivate proteins that require it for

part of their activity. This is very much in keeping with its type II topology, which makes it unlikely that these proteins transport heavy metals. There is strong support for this prediction in several patterns of genes co-expressed with the FUPA30 gene, such as the finding that multiple chemotaxis proteins, which require Ca^{2+} for proper function (Tisa LS, et al., 2000) are found to be coexpressed. Support for the involvement in other Ca^{2+}-mediated processes was also found. For example, several proteins co-expressed with FUPA30 protein are involved in cellular differention, such as sporulation in Bacilli (Wang SL, et al., 2008) and heterocyst formation in cyanobacteria (Hu Y, et al., 2011). In these processes, the Ca^{2+} concentration is an important determining factor. Protein folding, such as that mediated by DnaK- Ca^{2+}chaperone complexes, is catalyzed by many *dnaK/hsp* genes that are coexpressed with FUPA30 genes (Torrecilla I, et al., 2000; Norris V, et al., 1996). Changes in their activities may have far reaching effects on cellular activities. Support for this prediction also comes from the experimental confirmation that PacL, found both in *B.bacteriovorus* (a FUPA30 protein), and in *Synechococcus elongatus* PCC 7942, has been experimentally shown to be a Ca^{2+} transporter in *Synechococcus elongatus* PCC 7942 (Berkelman T, et al., 1994). Thus, based on its homology to the confirmed Ca^{2+} transporter PacL in *S.elongatus* PCC 7942, the FUPA30 of *B.bacteriovorus* may be predicted to function similarly.

FUPA31 is predicted to play a role in exopolysaccharide biosynthesis, cysteine biosynthesis, and/or inorganic sulfur assimilation. The predictions are made lightly, as the family possesses very few members available for study in SEED and none at all in

RegPredict. This being said, the only proteins found to be co-expressed with FUPA31 proteins were related to the above-stated functions.

Most of the FUPA proteins were found to be coexpressed with universal stress proteins UspA/E. For this reason, it is suggested that these genes are specialized, non-constitutive and essentially optional under homeostatic conditions. Possibly these enzymes allow a microbe to make “lifestyle choices” such as is necessary for non-obligate parasites. The relatively few members of these families compared to the members of known-function P-type ATPases, may be a testament to their nonessential nature in that they may have been eliminated more quickly, or only evolved in a small phylogenetic group of bacteria that are specialized for certain unique transport requirements. Also, several of the families with little or no co-expression with cation-requiring proteins may in fact be exporters, and thus not easily characterized by co-expression based on the cation needs of co-regulated proteins. Again, universal stress protein and perhaps even more importantly, heavy metal chaperone proteins, would be expected to be coexpressed with such heavy metal exporters. This provides a possible approach to the future characterization of such proteins.

The majority of the FUPAs are predicted to play roles in the transport of specific cations such as Ca for FUPA30 homologues, and Cu for FUPAs 27 and 29. This is in agreement with the specificities of known families they most closely resemble. The differences that set them apart phylogenetically, topologically and structurally are also what set them apart with respect to their proposed unique specialized role in the

timing/cell setting of their expression. Further work can be performed to confirm or refute these predictions. They serve as guides for further studies.

This report describes the first genome context analyses of this type conducted with the degree of breadth and depth reported here. It serves as a model to build upon in the field of functional analyses and predictions. Upon the introduction of more genomes containing members of the poorly populated FUPA families, this process can be used to improve predictions.

Table 1: Summary of functional predictions made for FUPA23-32

FUPA #	Proposed Functions
23	Mg^{2+} uptake transporter, providing the magnesium cation to proteins that require it for activity, these genes are predicted to be involved in cell division or antibiotic resistance
24	Mg^{2+}/divalent cation uptake transporter for proteins are predicted to be involved in biosynthesis of CoA, cytochrome c, lysyl-/isoleucyl-tRNAs
25	divalent cation uptake transporter, providing the cation to proteins that require it for activity, these genes are predicted to be involved in cell division regulation or cell division-associated protein expression, no strong predictions
26 & 32	$Fe^{2/3+}$/siderophore uptake transporter, providing the cation to proteins that require it for activity, these genes are predicted to be involved in A) heme biosynthesis and the cytochrome biogenesis pathway, or B) ferredoxin and ferrichrome activity/assembly, respectively
27 & 29	Cu^{2+} insert transporters, providing the cation to the cytochrome c oxidase complex (cco) or Nitrogen fixation (fix) proteins embedded in the membrane
28	either a Zn^{2+} or Fe^{2+} uptake transporter, harvesting the cations from within their hosts (Legionella-specific)
30	Ca^{2+} uptake transporter, providing the calcium cation to proteins, possibly by direct delivery, involved in chemotaxis, cellular differentiations such as sporulation and heterocyst formation, and chaperone folding protein activity modulation
31	role in exopolysaccharide or cysteine biosynthesis, or inorganic sulfur assimilation, no strong predictions

References

1. Ahn K, Kornberg A (1990) Polyphosphate kinase from Escherichia coli. Purification and demonstration of a phosphoenzyme intermediate. The Journal of biological chemistry 265: 11734-11739.

2. Airas RK (1996) Differences in the magnesium dependences of the class I and class II aminoacyl-tRNA synthetases from Escherichia coli. European journal of biochemistry / FEBS 240: 223-231.

3. Almog O, Kogan A, Leeuw M, Gdalevsky GY, Cohen-Luria R, et al. (2008) Structural insights into cold inactivation of tryptophanase and cold adaptation of subtilisin S41. Biopolymers 89: 354-359.

4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215: 403-410.

5. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389-3402.

6. Alwan AF, Mgbeje BI, Jordan PM (1989) Purification and properties of uroporphyrinogen III synthase (co-synthase) from an overproducing recombinant strain of Escherichia coli K-12. The Biochemical journal 264: 397-402.

7. Anand GS, Goudreau PN, Stock AM (1998) Activation of methylesterase CheB: evidence of a dual role for the regulatory domain. Biochemistry 37: 14038-14047.

8. Andorn N, Aronovitch J (1982) Purification and properties of histidinol dehydrogenase from Escherichia coli B. Journal of general microbiology 128: 579-584.

9. Andresson OS, Davies JE (1980) Some properties of the ribosomal RNA methyltransferase encoded by ksgA and the polarity of ksgA transcription. Molecular & general genetics : MGG 179: 217-222.

10. Arabshahi A, Flentke GR, Frey PA (1988) Uridine diphosphate galactose 4-epimerase. pH dependence of the reduction of NAD+ by a substrate analog. The Journal of biological chemistry 263: 2638-2643.

11. Arnoux P, Sabaty M, Alric J, Frangioni B, Guigliarelli B, et al. (2003) Structural and redox plasticity in the heterodimeric periplasmic nitrate reductase. Nature structural biology 10: 928-934.

12. Azzi A, Muller M (1990) Cytochrome c oxidases: polypeptide composition, role of subunits, and location of active metal centers. Archives of biochemistry and biophysics 280: 242-251.

13. Bailey AM, Mahapatra S, Brennan PJ, Crick DC (2002) Identification, cloning, purification, and enzymatic characterization of Mycobacterium tuberculosis 1-deoxy-D-xylulose 5-phosphate synthase. Glycobiology 12: 813-820.

14. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology 2: 28-36.

15. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48-54.

16. Ballantine SP, Volpe F, Delves CJ (1994) The hydroxymethyldihydropterin pyrophosphokinase domain of the multifunctional folic acid synthesis Fas protein of Pneumocystis carinii expressed as an independent enzyme in Escherichia coli: refolding and characterization of the recombinant enzyme. Protein expression and purification 5: 371-378.

17. Banks J, Cane DE (2004) Biosynthesis of vitamin B6: direct identification of the product of the PdxA-catalyzed oxidation of 4-hydroxy-l-threonine-4-phosphate using electrospray ionization mass spectrometry. Bioorganic & medicinal chemistry letters 14: 1633-1636.

18. Barrett AR, Kang Y, Inamasu KS, Son MS, Vukovich JM, et al. (2008) Genetic tools for allelic replacement in Burkholderia species. Applied and environmental microbiology 74: 4498-4508.

19. Benson BK, Meades G, Jr., Grove A, Waldrop GL (2008) DNA inhibits catalysis by the carboxyltransferase subunit of acetyl-CoA carboxylase: implications for active site communication. Protein science : a publication of the Protein Society 17: 34-42.

20. Berkelman T, Garret-Engele P, Hoffman NE (1994) The pacL gene of Synechococcus sp. strain PCC 7942 encodes a Ca(2+)-transporting ATPase. Journal of bacteriology 176: 4430-4436.

21. Berman T, Magasanik B (1966) The pathway of myo-inositol degradation in Aerobacter aerogenes. Dehydrogenation and dehydration. The Journal of biological chemistry 241: 800-806.

22. Bespalova IN, Burmeister M, Lesperance MM (1999) No association between DFNA6 and Pro250Arg mutation in FGFR3. American journal of medical genetics 88: 451.

23. Beyer W, Imlay J, Fridovich I (1991) Superoxide dismutases. Progress in nucleic acid research and molecular biology 40: 221-253.

24. Bhave DP, Hong JA, Lee M, Jiang W, Krebs C, et al. (2011) Spectroscopic studies on the [4Fe-4S] cluster in adenosine 5'-phosphosulfate reductase from Mycobacterium tuberculosis. The Journal of biological chemistry 286: 1216-1226.

25. Blair JM (1969) Magnesium and the aconitase equilibrium: determination of apparent stability constants of mangnesium substrate complexes from equilibrium data. European journal of biochemistry / FEBS 8: 287-291.

26. Blau N, Niederwieser A (1985) GTP-cyclohydrolases: a review. Journal of clinical chemistry and clinical biochemistry Zeitschrift fur klinische Chemie und klinische Biochemie 23: 169-176.

27. Boehlein SK, Richards NG, Schuster SM (1994) Glutamine-dependent nitrogen transfer in Escherichia coli asparagine synthetase B. Searching for the catalytic triad. The Journal of biological chemistry 269: 7450-7457.

28. Bognar AL, Shane B (1986) Bacterial folylpoly(gamma-glutamate) synthase-dihydrofolate synthase. Methods in enzymology 122: 349-359.

29. Bolesch DG, Keasling JD (2000) The effect of monovalent ions on polyphosphate binding to Escherichia coli exopolyphosphatase. Biochemical and biophysical research communications 274: 236-241.

30. Bollenbach TJ, Schuster G, Stern DB (2004) Cooperation of endo- and exoribonucleases in chloroplast mRNA turnover. Progress in nucleic acid research and molecular biology 78: 305-337.

31. Bond MD, Van Wart HE (1984) Characterization of the individual collagenases from Clostridium histolyticum. Biochemistry 23: 3085-3091.

32. Bouhss A, Crouvoisier M, Blanot D, Mengin-Lecreulx D (2004) Purification and characterization of the bacterial MraY translocase catalyzing the first membrane step of peptidoglycan biosynthesis. The Journal of biological chemistry 279: 29974-29980.

33. Brady DR, Houston LL (1973) Some properties of the catalytic sites of imidazoleglycerol phosphate dehydratase-histidinol phosphate phosphatase, a bifunctional enzyme from Salmonella typhimurium. The Journal of biological chemistry 248: 2588-2592.

34. Brand BC, Sadosky AB, Shuman HA (1994) The Legionella pneumophila icm locus: a set of genes required for intracellular multiplication in human macrophages. Molecular microbiology 14: 797-808.

35. Breckau D, Mahlitz E, Sauerwald A, Layer G, Jahn D (2003) Oxygen-dependent coproporphyrinogen III oxidase (HemF) from Escherichia coli is stimulated by manganese. The Journal of biological chemistry 278: 46625-46631.

36. Brot N, Weissbach L, Werth J, Weissbach H (1981) Enzymatic reduction of protein-bound methionine sulfoxide. Proceedings of the National Academy of Sciences of the United States of America 78: 2155-2158.

37. Buchko GW, Hess NJ, Bandaru V, Wallace SS, Kennedy MA (2000) Spectroscopic studies of zinc(II)- and cobalt(II)-associated Escherichia coli formamidopyrimidine-DNA glycosylase: extended X-ray absorption fine structure evidence for a metal-binding domain. Biochemistry 39: 12441-12449.

38. Busch W, Saier MH, Jr. (2002) The transporter classification (TC) system, 2002. Critical reviews in biochemistry and molecular biology 37: 287-337.

39. Campos-Bermudez VA, Moran-Barrio J, Costa-Filho AJ, Vila AJ (2010) Metal-dependent inhibition of glyoxalase II: a possible mechanism to regulate the enzyme activity. Journal of inorganic biochemistry 104: 726-731.

40. Castillo R, Saier MH (2010) Functional Promiscuity of Homologues of the Bacterial ArsA ATPases. International journal of microbiology 2010: 187373.

41. Chan SI, Li PM (1990) Cytochrome c oxidase: understanding nature's design of a proton pump. Biochemistry 29: 1-12.

42. Christensen KE, Mackenzie RE (2008) Mitochondrial methylenetetrahydrofolate dehydrogenase, methenyltetrahydrofolate cyclohydrolase, and formyltetrahydrofolate synthetases. Vitamins and hormones 79: 393-410.

43. Cohen SS (1951) Gluconokinase and the oxidative path of glucose-6-phosphate utilization. The Journal of biological chemistry 189: 617-628.

44. Conlin CA, Hakensson K, Liljas A, Miller CG (1994) Cloning and nucleotide sequence of the cyclic AMP receptor protein-regulated Salmonella typhimurium pepE gene and crystallization of its product, an alpha-aspartyl dipeptidase. Journal of bacteriology 176: 166-172.

45. Crepin T, Yaremchuk A, Tukalo M, Cusack S (2006) Structures of two bacterial prolyl-tRNA synthetases with and without a cis-editing domain. Structure 14: 1511-1525.

46. Curley P, van der Does C, Driessen AJ, Kok J, van Sinderen D (2003) Purification and characterisation of a lactococcal aminoacylase. Archives of microbiology 179: 402-408.

47. Davidson AL, Dassa E, Orelle C, Chen J (2008) Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiology and molecular biology reviews : MMBR 72: 317-364, table of contents.

48. de Carvalho LP, Blanchard JS (2006) Kinetic analysis of the effects of monovalent cations and divalent metals on the activity of Mycobacterium tuberculosis alpha-isopropylmalate synthase. Archives of biochemistry and biophysics 451: 141-148.

49. Del Papa MF, Perego M (2008) Ethanolamine activates a sensor histidine kinase regulating its utilization in Enterococcus faecalis. Journal of bacteriology 190: 7147-7156.

50. Delafield FP, Cooksey KE, Doudoroff M (1965) beta-Hydroxybutyric dehydrogenase and dimer hydrolase of Pseudomonas lemoignei. The Journal of biological chemistry 240: 4023-4028.

51. Deobagkar DN, Gopinathan KP (1976) Two forms of methionyl-transfer RNA synthetase from Mycobacterium smegmatis. Biochemical and biophysical research communications 71: 939-951.

52. Dhiman RK, Schulbach MC, Mahapatra S, Baulard AR, Vissa V, et al. (2004) Identification of a novel class of omega,E,E-farnesyl diphosphate synthase from Mycobacterium tuberculosis. Journal of lipid research 45: 1140-1147.

53. Dierkers AT, Niks D, Schlichting I, Dunn MF (2009) Tryptophan synthase: structure and function of the monovalent cation site. Biochemistry 48: 10997-11010.

54. Dorner E, Boll M (2002) Properties of 2-oxoglutarate:ferredoxin oxidoreductase from Thauera aromatica and its role in enzymatic reduction of the aromatic ring. Journal of bacteriology 184: 3975-3983.

55. Duggleby RG, McCourt JA, Guddat LW (2008) Structure and mechanism of inhibition of plant acetohydroxyacid synthase. Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale 46: 309-324.

56. Dumas R, Biou V, Halgand F, Douce R, Duggleby RG (2001) Enzymology, structure, and dynamics of acetohydroxy acid isomeroreductase. Accounts of chemical research 34: 399-408.

57. Dungan SM, Katta P (1973) Concerted feedback inhibition. Purification and some properties of aspartokinase from Pseudomonas fluorescens. The Journal of biological chemistry 248: 8534-8540.

58. Eady RR, Postgate JR (1974) Nitrogenase. Nature 249: 805-810.

59. Egan AF, Gibson F (1972) Anthranilate synthase-anthranilate 5-phosphoribosyl 1-pyrophosphate phosphoribosyltransferase from Aerobacter aerogenes. The Biochemical journal 130: 847-859.

60. Elleby B, Chirica LC, Tu C, Zeppezauer M, Lindskog S (2001) Characterization of carbonic anhydrase from Neisseria gonorrhoeae. European journal of biochemistry / FEBS 268: 1613-1619.

61. Erb TJ, Ismail W, Fuchs G (2008) Phenylacetate metabolism in thermophiles: characterization of phenylacetate-CoA ligase, the initial enzyme of the hybrid pathway in Thermus thermophilus. Current microbiology 57: 27-32.

62. Erwin AL, Gotschlich EC (1993) Oxidation of D-lactate and L-lactate by Neisseria meningitidis: purification and cloning of meningococcal D-lactate dehydrogenase. Journal of bacteriology 175: 6382-6391.

63. Flemming D, Hellwig P, Lepper S, Kloer DP, Friedrich T (2006) Catalytic importance of acidic amino acids on subunit NuoB of the Escherichia coli NADH:ubiquinone oxidoreductase (complex I). The Journal of biological chemistry 281: 24781-24789.

64. Flemming D, Stolpe S, Schneider D, Hellwig P, Friedrich T (2005) A possible role for iron-sulfur cluster N2 in proton translocation by the NADH: ubiquinone oxidoreductase (complex I). Journal of molecular microbiology and biotechnology 10: 208-222.

65. Foot M, Jeffcoat R, Russell NJ (1983) Some properties, including the substrate in vivo, of the delta 9-desaturase in Micrococcus cryophilus. The Biochemical journal 209: 345-353.

66. Fordyce AM, Moore CH, Pritchard GG (1982) Phosphofructokinase from Streptococcus lactis. Methods in enzymology 90 Pt E: 77-82.

67. Forouhar F, Hussain M, Farid R, Benach J, Abashidze M, et al. (2006) Crystal structures of two bacterial 3-hydroxy-3-methylglutaryl-CoA lyases suggest a common catalytic mechanism among a family of TIM barrel metalloenzymes cleaving carbon-carbon bonds. The Journal of biological chemistry 281: 7533-7545.

68. Friedrich P, Darley DJ, Golding BT, Buckel W (2008) The complete stereochemistry of the enzymatic dehydration of 4-hydroxybutyryl coenzyme A to crotonyl coenzyme A. Angewandte Chemie 47: 3254-3257.

69. Furci LM, Lopes P, Eakanunkul S, Zhong S, MacKerell AD, Jr., et al. (2007) Inhibition of the bacterial heme oxygenases from Pseudomonas aeruginosa and Neisseria meningitidis: novel antimicrobial targets. Journal of medicinal chemistry 50: 3804-3813.

70. Geissler JF, Harwood CS, Gibson J (1988) Purification and properties of benzoate-coenzyme A ligase, a Rhodopseudomonas palustris enzyme involved in the anaerobic degradation of benzoate. Journal of bacteriology 170: 1709-1714.

71. Gill HS, Pfluegl GM, Eisenberg D (2002) Multicopy crystallographic refinement of a relaxed glutamine synthetase from Mycobacterium tuberculosis highlights flexible loops in the enzymatic mechanism and its regulation. Biochemistry 41: 9863-9872.

72. Gong H, Murphy A, McMaster CR, Byers DM (2007) Neutralization of acidic residues in helix II stabilizes the folded conformation of acyl carrier protein and variably alters its function with different enzymes. The Journal of biological chemistry 282: 4494-4503.

73. Gopalaswamy R, Narayanan PR, Narayanan S (2004) Cloning, overexpression, and characterization of a serine/threonine protein kinase pknI from Mycobacterium tuberculosis H37Rv. Protein expression and purification 36: 82-89.

74. Gough SP, Rzeznicka K, Peterson Wulff R, Francisco Jda C, Hansson A, et al. (2007) A new method for isolating physiologically active Mg-protoporphyrin monomethyl ester, the substrate of the cyclase enzyme of the chlorophyll biosynthetic pathway. Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale 45: 932-936.

75. Goward CR, Hartwell R, Atkinson T, Scawen MD (1986) The purification and characterization of glucokinase from the thermophile Bacillus stearothermophilus. The Biochemical journal 237: 415-420.

76. Grawert T, Kaiser J, Zepeck F, Laupitz R, Hecht S, et al. (2004) IspH protein of Escherichia coli: studies on iron-sulfur cluster implementation and catalysis. Journal of the American Chemical Society 126: 12847-12855.

77. Grimek TL, Escalante-Semerena JC (2004) The acnD genes of Shewenella oneidensis and Vibrio cholerae encode a new Fe/S-dependent 2-methylcitrate dehydratase enzyme that requires prpF function in vivo. Journal of bacteriology 186: 454-462.

78. Guchhait RB, Polakis SE, Dimroth P, Stoll E, Moss J, et al. (1974) Acetyl coenzyme A carboxylase system of Escherichia coli. Purification and properties of the biotin carboxylase, carboxyltransferase, and carboxyl carrier protein components. The Journal of biological chemistry 249: 6633-6645.

79. Guengerich FP, Fang Q, Liu L, Hachey DL, Pegg AE (2003) O6-alkylguanine-DNA alkyltransferase: low pKa and high reactivity of cysteine 145. Biochemistry 42: 10965-10970.

80. Gushima H, Miya T, Murata K, Kimura A (1983) Purification and characterization of glutathione synthetase from Escherichia coli B. Journal of applied biochemistry 5: 210-218.

81. Haas W, Sublett J, Kaushal D, Tuomanen EI (2004) Revising the role of the pneumococcal vex-vncRS locus in vancomycin tolerance. Journal of bacteriology 186: 8463-8471.

82. Hagen AR, Barabote RD, Saier MH (2005) The bestrophin family of anion channels: identification of prokaryotic homologues. Molecular membrane biology 22: 291-302.

83. Hasan N, Nester EW (1978) Purification and properties of chorismate synthase from Bacillus subtilis. The Journal of biological chemistry 253: 4993-4998.

84. Hassani BK, Astier C, Nitschke W, Ouchane S (2010) CtpA, a copper-translocating P-type ATPase involved in the biogenesis of multiple copper-requiring enzymes. The Journal of biological chemistry 285: 19330-19337.

85. Havukainen H, Haataja S, Kauko A, Pulliainen AT, Salminen A, et al. (2008) Structural basis of the zinc- and terbium-mediated inhibition of ferroxidase activity in Dps ferritin-like proteins. Protein science : a publication of the Protein Society 17: 1513-1521.

86. Hayzer DJ, Leisinger T (1982) Proline biosynthesis in Escherichia coli. Purification and characterisation of glutamate-semialdehyde dehydrogenase. European journal of biochemistry / FEBS 121: 561-565.

87. Heinrikson RL, Goldwasser E (1964) Studies on the Biosynthesis of 5-Ribosyluracil 5'-Monophosphate in Tetrahymena Pyriformis. The Journal of biological chemistry 239: 1177-1187.

88. Hele P, Barber R (1972) Lysyl tRNA synthetase of Escherichia coli B: formation and reactions of ATP-enzyme and lysyl-AMP-enzyme complexes. Biochimica et biophysica acta 258: 319-331.

89. Hensel R, Mayr U, Stetter KO, Kandler O (1977) Comparative studies of lactic acid dehydrogenases in lactic acid bacteria. I. Purification and kinetics of the allosteric L-lactic acid dehydrogenase from Lactobacillus casei ssp. casei and Lactobacillus curvatus. Archives of microbiology 112: 81-93.

90. Hesketh A, Kock H, Mootien S, Bibb M (2009) The role of absC, a novel regulatory gene for secondary metabolism, in zinc-dependent antibiotic production in Streptomyces coelicolor A3(2). Molecular microbiology 74: 1427-1444.

91. Hindra, Pak P, Elliot MA (2010) Regulation of a novel gene cluster involved in secondary metabolite production in Streptomyces coelicolor. Journal of bacteriology 192: 4973-4982.

92. Hochstadt J (1978) Adenine phosphoribosyltransferase from Escherichia coli. Methods in enzymology 51: 558-567.

93. Hong BS, Yun MK, Zhang YM, Chohnan S, Rock CO, et al. (2006) Prokaryotic type II and type III pantothenate kinases: The same monomer fold creates dimers with distinct catalytic properties. Structure 14: 1251-1261.

94. Hosaka T, Meguro T, Yamato I, Shirakihara Y (2003) Crystal structure of Enterococcus hirae enolase at 2.8 A resolution. Journal of biochemistry 133: 817-823.

95. Hoving H, Koning JH, Robillard GT (1982) Escherichia coli phosphoenolpyruvate-dependent phosphotransferase system: role of divalent metals in the dimerization and phosphorylation of enzyme I. Biochemistry 21: 3128-3136.

96. Hsu JL, Chen HC, Peng HL, Chang HY (2008) Characterization of the histidine-containing phosphotransfer protein B-mediated multistep phosphorelay system in Pseudomonas aeruginosa PAO1. The Journal of biological chemistry 283: 9933-9944.

97. Hu Y, Zhang X, Shi Y, Zhou Y, Zhang W, et al. (2011) Structures of Anabaena calcium-binding protein CcbP: insights into Ca2+ signaling during heterocyst differentiation. The Journal of biological chemistry 286: 12381-12388.

98. Husic HD, Tolbert NE (1985) Properties of Phosphoglycolate Phosphatase from Chlamydomonas reinhardtii and Anacystis nidulans. Plant physiology 79: 394-399.

99. Iding H, Dunnwald T, Greiner L, Liese A, Muller M, et al. (2000) Benzoylformate decarboxylase from Pseudomonas putida as stable catalyst for the synthesis of chiral 2-hydroxy ketones. Chemistry 6: 1483-1495.

100. Imburgio D, Anikin M, McAllister WT (2002) Effects of substitutions in a conserved DX(2)GR sequence motif, found in many DNA-dependent nucleotide polymerases, on transcription by T7 RNA polymerase. Journal of molecular biology 319: 37-51.

101. Imsande J (1961) Pathway of diphosphopyridine nucleotide biosynthesis in Escherichia coli. The Journal of biological chemistry 236: 1494-1497.

102. Itami H, Sakai Y, Shimamoto T, Hama H, Tsuda M, et al. (1989) Purification and characterization of membrane-bound 5'-nucleotidase of Vibrio parahaemolyticus. Journal of biochemistry 105: 785-789.

103. Ito K, Nakajima Y, Onohara Y, Takeo M, Nakashima K, et al. (2006) Crystal structure of aminopeptidase N (proteobacteria alanyl aminopeptidase) from Escherichia coli and conformational change of methionine 260 involved in substrate recognition. The Journal of biological chemistry 281: 33664-33676.

104. Jacobs NJ, Borotz SE, Jacobs JM (1989) Characteristics of purified protoporphyrinogen oxidase from barley. Biochemical and biophysical research communications 161: 790-796.

105. Javid-Majd F, Blanchard JS (2000) Mechanistic analysis of the argE-encoded N-acetylornithine deacetylase. Biochemistry 39: 1285-1293.

106. Javid-Majd F, Yang D, Ioerger TR, Sacchettini JC (2008) The 1.25 A resolution structure of phosphoribosyl-ATP pyrophosphohydrolase from Mycobacterium tuberculosis. Acta crystallographica Section D, Biological crystallography 64: 627-635.

107. Jeyakanthan J, Thamotharan S, Velmurugan D, Rao VS, Nagarajan S, et al. (2009) New structural insights and molecular-modelling studies of 4-methyl-5-beta-hydroxyethylthiazole kinase from Pyrococcus horikoshii OT3 (PhThiK). Acta crystallographica Section F, Structural biology and crystallization communications 65: 978-986.

108. Johnson CL, Buszko ML, Bobik TA (2004) Purification and initial characterization of the Salmonella enterica PduO ATP:Cob(I)alamin adenosyltransferase. Journal of bacteriology 186: 7881-7887.

109. Jones RM, Jordan PM (1993) Purification and properties of the uroporphyrinogen decarboxylase from Rhodobacter sphaeroides. The Biochemical journal 293 (Pt 3): 703-712.

110. Jordens J, Janssens V, Longin S, Stevens I, Martens E, et al. (2006) The protein phosphatase 2A phosphatase activator is a novel peptidyl-prolyl cis/trans-isomerase. The Journal of biological chemistry 281: 6349-6357.

111. Joyce MA, Fraser ME, James MN, Bridger WA, Wolodko WT (2000) ADP-binding site of Escherichia coli succinyl-CoA synthetase revealed by x-ray crystallography. Biochemistry 39: 17-25.

112. Kahn D, David M, Domergue O, Daveran ML, Ghai J, et al. (1989) Rhizobium meliloti fixGHI sequence predicts involvement of a specific cation pump in symbiotic nitrogen fixation. Journal of bacteriology 171: 929-939.

113. Kaiser J, Schramek N, Eberhardt S, Puttmer S, Schuster M, et al. (2002) Biosynthesis of vitamin B2. European journal of biochemistry / FEBS 269: 5264-5270.

114. Kang YN, Tran A, White RH, Ealick SE (2007) A novel function for the N-terminal nucleophile hydrolase fold demonstrated by the structure of an archaeal inosine monophosphate cyclohydrolase. Biochemistry 46: 5050-5062.

115. Kapoor R, Venkitasubramanian TA (1983) Purification and properties of pyruvate kinase from Mycobacterium smegmatis. Archives of biochemistry and biophysics 225: 320-330.

116. Kawai F, Kimura T, Tani Y, Yamada H, Kurachi M (1980) Purification and characterization of polyethylene glycol dehydrogenase involved in the bacterial metabolism of polyethylene glycol. Applied and environmental microbiology 40: 701-705.

117. Kawai S, Murata K (2008) Structure and function of NAD kinase and NADP phosphatase: key enzymes that regulate the intracellular balance of NAD(H) and NADP(H). Bioscience, biotechnology, and biochemistry 72: 919-930.

118. Kawasaki T (1979) Thiamine phosphate pyrophosphorylase. Methods in enzymology 62: 69-73.

119. Kayama Y, Kawasaki T (1973) Purification and properties of thiaminephosphate pyrophosphorylase of Escherichia coli. Archives of biochemistry and biophysics 158: 242-248.

120. Kelly BS, Antholine WE, Griffith OW (2002) Escherichia coli gamma-glutamylcysteine synthetase. Two active site metal ions affect substrate and inhibitor binding. The Journal of biological chemistry 277: 50-58.

121. Kim S (1984) S-adenosylmethionine: protein-carboxyl O-methyltransferase (protein methylase II). Methods in enzymology 106: 295-309.

122. Kim S, Jo YJ, Lee SH, Motegi H, Shiba K, et al. (1998) Biochemical and phylogenetic analyses of methionyl-tRNA synthetase isolated from a pathogenic microorganism, Mycobacterium tuberculosis. FEBS letters 427: 259-262.

123. Kim SA, Copeland L (1997) Acetyl Coenzyme A Acetyltransferase of Rhizobium sp. (Cicer) Strain CC 1192. Applied and environmental microbiology 63: 3432-3437.

124. Kimura Y, Okazaki N, Takegawa K (2009) Enzymatic characteristics of two novel Myxococcus xanthus enzymes, PdeA and PdeB, displaying 3',5'- and 2',3'-cAMP phosphodiesterase, and phosphatase activities. FEBS letters 583: 443-448.

125. Kiyasu T, Asakura A, Nagahashi Y, Hoshino T (2002) Biotin synthase of Bacillus subtilis shows less reactivity than that of Escherichia coli in in vitro reaction systems. Archives of microbiology 179: 26-32.

126. Koreishi M, Nakatani Y, Ooi M, Imanaka H, Imamura K, et al. (2009) Purification, characterization, molecular cloning, and expression of a new aminoacylase from Streptomyces mobaraensis that can hydrolyze N-(middle/long)-chain-fatty-acyl-

L-amino acids as well as N-short-chain-acyl-L-amino acids. Bioscience, biotechnology, and biochemistry 73: 1940-1947.

127. Kornberg T, Gefter ML (1974) Deoxyribonucleic acid polymerase 3 (Escherichia coli K12). Methods in enzymology 29: 22-26.

128. Krebs A, Bridger WA (1976) On the monomeric structure and proposed regulatory properties of phosphoenolpyruvate carboxykinase of Escherichia coli. Canadian journal of biochemistry 54: 22-26.

129. Krebs A, Bridger WA (1980) The kinetic properties of phosphoenolpyruvate carboxykinase of Escherichia coli. Canadian journal of biochemistry 58: 309-318.

130. Kuhn NJ, Setlow B, Setlow P (1993) Manganese(II) activation of 3-phosphoglycerate mutase of Bacillus megaterium: pH-sensitive interconversion of active and inactive forms. Archives of biochemistry and biophysics 306: 342-349.

131. Kumar P, Singh M, Gautam R, Karthikeyan S (2010) Potential anti-bacterial drug target: structural characterization of 3,4-dihydroxy-2-butanone-4-phosphate synthase from Salmonella typhimurium LT2. Proteins 78: 3292-3303.

132. Kyrtopoulos SA, Satchell DP (1973) Kinetic studies with phosphotransacetylase. V. The mechanism of activation by univalent cations. Biochimica et biophysica acta 321: 126-142.

133. Lacher K, Schafer G (1993) Archaebacterial adenylate kinase from the thermoacidophile Sulfolobus acidocaldarius: purification, characterization, and partial sequence. Archives of biochemistry and biophysics 302: 391-397.

134. Law AS, Wriston JC, Jr. (1971) Purification and properties of Bacillus coagulans L-asparaginase. Archives of biochemistry and biophysics 147: 744-752.

135. Layer G, Kervio E, Morlock G, Heinz DW, Jahn D, et al. (2005) Structural and functional comparison of HemN to other radical SAM enzymes. Biological chemistry 386: 971-980.

136. Le KH, Boussac A, Frangioni B, Leger C, Lederer F (2009) Interdomain contacts in flavocytochrome b(2), a mutational analysis. Biochemistry 48: 10803-10809.

137. Lehmann M, Tshisuaka B, Fetzner S, Roger P, Lingens F (1994) Purification and characterization of isoquinoline 1-oxidoreductase from Pseudomonas diminuta 7, a novel molybdenum-containing hydroxylase. The Journal of biological chemistry 269: 11254-11260.

138. Levy CC, Goldman P (1970) Residue specificity of a ribonuclease which hydrolyzes polycytidylic acid. The Journal of biological chemistry 245: 3257-3262.

139. Leyva LA, Bashan Y (2008) Activity of two catabolic enzymes of the phosphogluconate pathway in mesquite roots inoculated with Azospirillum brasilense Cd. Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale 46: 898-904.

140. Liu J, Lin SX, Blochet JE, Pezolet M, Lapointe J (1993) The glutamyl-tRNA synthetase of Escherichia coli contains one atom of zinc essential for its native conformation and its catalytic activity. Biochemistry 32: 11390-11396.

141. Liu W, Peterson PE, Carter RJ, Zhou X, Langston JA, et al. (2004) Crystal structures of unbound and aminooxyacetate-bound Escherichia coli gamma-aminobutyrate aminotransferase. Biochemistry 43: 10896-10905.

142. Liu WT, Karavolos MH, Bulmer DM, Allaoui A, Hormaeche RD, et al. (2007) Role of the universal stress protein UspA of Salmonella in growth arrest, stress and virulence. Microbial pathogenesis 42: 2-10.

143. Lohkamp B, McDermott G, Campbell SA, Coggins JR, Lapthorn AJ (2004) The structure of Escherichia coli ATP-phosphoribosyltransferase: identification of substrate binding sites and mode of AMP inhibition. Journal of molecular biology 336: 131-144.

144. Ludwig B (1980) Heme aa3-type cytochrome c oxidases from bacteria. Biochimica et biophysica acta 594: 177-189.

145. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic acids research 37: D205-210.

146. Marco-Marin C, Gil-Ortiz F, Perez-Arellano I, Cervera J, Fita I, et al. (2007) A novel two-domain architecture within the amino acid kinase enzyme family revealed by the crystal structure of Escherichia coli glutamate 5-kinase. Journal of molecular biology 367: 1431-1446.

147. Markham GD, Hafner EW, Tabor CW, Tabor H (1980) S-Adenosylmethionine synthetase from Escherichia coli. The Journal of biological chemistry 255: 9082-9092.

148. Marrakchi H, Dewolf WE, Jr., Quinn C, West J, Polizzi BJ, et al. (2003) Characterization of Streptococcus pneumoniae enoyl-(acyl-carrier protein) reductase (FabK). The Biochemical journal 370: 1055-1062.

149. Matak-Vinkovic D, Vinkovic M, Saldanha SA, Ashurst JL, von Delft F, et al. (2001) Crystal structure of Escherichia coli ketopantoate reductase at 1.7 A resolution and insight into the enzyme mechanism. Biochemistry 40: 14493-14500.

150. Mathew Z, Knox TM, Miller CG (2000) Salmonella enterica serovar typhimurium peptidase B is a leucyl aminopeptidase with specificity for acidic amino acids. Journal of bacteriology 182: 3383-3393.

151. Mathur D, Malik G, Garg LC (2006) Biochemical and functional characterization of triosephosphate isomerase from Mycobacterium tuberculosis H37Rv. FEMS microbiology letters 263: 229-235.

152. Mayer SM, Rieble S, Beale SI (1994) Metal requirements of the enzymes catalyzing conversion of glutamate to delta-aminolevulinic acid in extracts of Chlorella vulgaris and Synechocystis sp. PCC 6803. Archives of biochemistry and biophysics 312: 203-209.

153. McCutcheon JP, McDonald BR, Moran NA (2009) Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proceedings of the National Academy of Sciences of the United States of America 106: 15394-15399.

154. McRee DE, Richardson DC, Richardson JS, Siegel LM (1986) The heme and Fe4S4 cluster in the crystallographic structure of Escherichia coli sulfite reductase. The Journal of biological chemistry 261: 10277-10281.

155. Mijakovic I, Musumeci L, Tautz L, Petranovic D, Edwards RA, et al. (2005) In vitro characterization of the Bacillus subtilis protein tyrosine phosphatase YwqE. Journal of bacteriology 187: 3384-3390.

156. Mikolajek R, Spiess AC, Buchs J (2007) Feasibility of gas/solid carboligation: conversion of benzaldehyde to benzoin using thiamine diphosphate-dependent enzymes. Journal of biotechnology 129: 723-725.

157. Miller RE, Stadtman ER (1972) Glutamate synthase from Escherichia coli. An iron-sulfide flavoprotein. The Journal of biological chemistry 247: 7407-7419.

158. Mustafi D, Bekesi A, Vertessy BG, Makinen MW (2003) Catalytic and structural role of the metal ion in dUTP pyrophosphatase. Proceedings of the National Academy of Sciences of the United States of America 100: 5670-5675.

159. Myers JW (1961) Dihydroxy acid dehydrase: an enzyme involved in the biosynthesis of isoleucine and valine. The Journal of biological chemistry 236: 1414-1418.

160. Nakagawa J, Tamaki S, Tomioka S, Matsuhashi M (1984) Functional biosynthesis of cell wall peptidoglycan by polymorphic bifunctional polypeptides. Penicillin-binding protein 1Bs of Escherichia coli with activities of transglycosylase and transpeptidase. The Journal of biological chemistry 259: 13937-13946.

161. Neidhart DJ, Kenyon GL, Gerlt JA, Petsko GA (1990) Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. Nature 347: 692-694.

162. Nelson DJ, Carter CE (1969) Purification and characterization of Thymidine 5-monophosphate kinase from Escherichia coli B. The Journal of biological chemistry 244: 5254-5262.

163. Neudert U, Martinez-Ferez IM, Fraser PD, Sandmann G (1998) Expression of an active phytoene synthase from Erwinia uredovora and biochemical properties of the enzyme. Biochimica et biophysica acta 1392: 51-58.

164. Nilekani S, SivaRaman C (1983) Purification and properties of citrate lyase from Escherichia coli. Biochemistry 22: 4657-4663.

165. Ning B, Elbein AD (1999) Purification and properties of mycobacterial GDP-mannose pyrophosphorylase. Archives of biochemistry and biophysics 362: 339-345.

166. Nixon PF, Blakley RL (1968) Dihydrofolate reductase of Streptococcus faicium. II. Purification and some properties of two dihydrofolate reductases from the amethopterin-resistant mutant Streptococcus faecium var. Durans strain A. The Journal of biological chemistry 243: 4722-4731.

167. Norris V, Grant S, Freestone P, Canvin J, Sheikh FN, et al. (1996) Calcium signalling in bacteria. Journal of bacteriology 178: 3677-3682.

168. Novichkov PS, Rodionov DA, Stavrovskaya ED, Novichkova ES, Kazakov AE, et al. (2010) RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. Nucleic acids research 38: W299-307.

169. Nystrom T, Neidhardt FC (1994) Expression and role of the universal stress protein, UspA, of Escherichia coli during growth arrest. Molecular microbiology 11: 537-544.

170. Ogasawara Y, Liu HW (2009) Biosynthetic studies of aziridine formation in azicemicins. Journal of the American Chemical Society 131: 18066-18068.

171. Ogata H, Stolle P, Stehr M, Auling G, Lubitz W (2009) Crystallization and preliminary X-ray analysis of the small subunit (R2F) of native ribonucleotide reductase from Corynebacterium ammoniagenes. Acta crystallographica Section F, Structural biology and crystallization communications 65: 878-880.

172. Ohtani N, Haruki M, Muroya A, Morikawa M, Kanaya S (2000) Characterization of ribonuclease HII from Escherichia coli overproduced in a soluble form. Journal of biochemistry 127: 895-899.

173. Okada K, Hase T (2005) Cyanobacterial non-mevalonate pathway: (E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase interacts with ferredoxin in Thermosynechococcus elongatus BP-1. The Journal of biological chemistry 280: 20672-20679.

174. Okuyama H, Wakil SJ (1973) Positional specificities of acyl coenzyme A: glycerophosphate and acyl coenzyme A: monoacylglycerophosphate acyltransferases in Escherichia coli. The Journal of biological chemistry 248: 5197-5205.

175. Olry A, Boschi-Muller S, Yu H, Burnel D, Branlant G (2005) Insights into the role of the metal binding site in methionine-R-sulfoxide reductases B. Protein science : a publication of the Protein Society 14: 2828-2837.

176. Omi R, Goto M, Miyahara I, Manzoku M, Ebihara A, et al. (2007) Crystal structure of monofunctional histidinol phosphate phosphatase from Thermus thermophilus HB8. Biochemistry 46: 12618-12627.

177. Orengo A, Kobayashi SH (1978) Uridine-cytidine kinase from Novikoff ascites rat tumor and Bacillus stearothermophilus. Methods in enzymology 51: 299-307.

178. Orlowski J, Bujnicki JM (2008) Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses. Nucleic acids research 36: 3552-3569.

179. Outten FW, Djaman O, Storz G (2004) A suf operon requirement for Fe-S cluster assembly during iron starvation in Escherichia coli. Molecular microbiology 52: 861-872.

180. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic acids research 33: 5691-5702.

181. Ozment C, Barchue J, DeLucas LJ, Chattopadhyay D (1999) Structural study of Escherichia coli NAD synthetase: overexpression, purification, crystallization, and preliminary crystallographic analysis. Journal of structural biology 127: 279-282.

182. Pacaud M, Richaud C (1975) Protease II from Escherichia coli. Purification and characterization. The Journal of biological chemistry 250: 7771-7779.

183. Park CS, Yeom SJ, Kim HJ, Lee SH, Lee JK, et al. (2007) Characterization of ribose-5-phosphate isomerase of Clostridium thermocellum producing D-allose from D-psicose. Biotechnology letters 29: 1387-1391.

184. Parsons JF, Jensen PY, Pachikara AS, Howard AJ, Eisenstein E, et al. (2002) Structure of Escherichia coli aminodeoxychorismate synthase: architectural conservation and diversity in chorismate-utilizing enzymes. Biochemistry 41: 2198-2208.

185. Pasman Z, Robey-Bond S, Mirando AC, Smith GJ, Lague A, et al. (2011) Substrate specificity and catalysis by the editing active site of Alanyl-tRNA synthetase from Escherichia coli. Biochemistry 50: 1474-1482.

186. Patra B, Ghosh Dastidar K, Maitra S, Bhattacharyya J, Majumder AL (2007) Functional identification of sll1383 from Synechocystis sp PCC 6803 as L-myo-inositol 1-phosphate phosphatase (EC 3.1.3.25): molecular cloning, expression and characterization. Planta 225: 1547-1558.

187. Peters-Wendisch P, Stolz M, Etterich H, Kennerknecht N, Sahm H, et al. (2005) Metabolic engineering of Corynebacterium glutamicum for L-serine production. Applied and environmental microbiology 71: 7139-7144.

188. Pierce BS, Hendrich MP (2005) Local and global effects of metal binding within the small subunit of ribonucleotide reductase. Journal of the American Chemical Society 127: 3613-3623.

189. Porankiewicz J, Wang J, Clarke AK (1999) New insights into the ATP-dependent Clp protease: Escherichia coli and beyond. Molecular microbiology 32: 449-458.

190. Porter TN, Li Y, Raushel FM (2004) Mechanism of the dihydroorotase reaction. Biochemistry 43: 16285-16292.

191. Powers SG, Snell EE (1976) Ketopantoate hydroxymethyltransferase. II. Physical, catalytic, and regulatory properties. The Journal of biological chemistry 251: 3786-3793.

192. Powers SG, Snell EE (1979) Purification and properties of ketopantoate hydroxymethyltransferase. Methods in enzymology 62: 204-209.

193. Pratviel-Sosa F, Mengin-Lecreulx D, van Heijenoort J (1991) Over-production, purification and properties of the uridine diphosphate N-acetylmuramoyl-L-alanine:D-glutamate ligase from Escherichia coli. European journal of biochemistry / FEBS 202: 1169-1176.

194. Preisig O, Zufferey R, Hennecke H (1996) The Bradyrhizobium japonicum fixGHIS genes are required for the formation of the high-affinity cbb3-type cytochrome oxidase. Archives of microbiology 165: 297-305.

195. Preiss J (1978) Regulation of adenosine diphosphate glucose pyrophosphorylase. Advances in enzymology and related areas of molecular biology 46: 317-381.

196. Preston GG, Wall JD, Emerich DW (1990) Purification and properties of acetyl-CoA synthetase from Bradyrhizobium japonicum bacteroids. The Biochemical journal 267: 179-183.

197. Price AC, Rock CO, White SW (2003) The 1.3-Angstrom-resolution crystal structure of beta-ketoacyl-acyl carrier protein synthase II from Streptococcus pneumoniae. Journal of bacteriology 185: 4136-4143.

198. Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic acids research 33: 880-892.

199. Proudfoot M, Kuznetsova E, Brown G, Rao NN, Kitagawa M, et al. (2004) General enzymatic screens identify three new nucleotidases in Escherichia coli.

Biochemical characterization of SurE, YfbR, and YjjG. The Journal of biological chemistry 279: 54687-54694.

200. Proudfoot M, Sanders SA, Singer A, Zhang R, Brown G, et al. (2008) Biochemical and structural characterization of a novel family of cystathionine beta-synthase domain proteins fused to a Zn ribbon-like domain. Journal of molecular biology 375: 301-315.

201. Rajagopalan PT, Pei D (1998) Oxygen-mediated inactivation of peptide deformylase. The Journal of biological chemistry 273: 22305-22310.

202. Ramsden NL, Buetow L, Dawson A, Kemp LA, Ulaganathan V, et al. (2009) A structure-based approach to ligand discovery for 2C-methyl-D-erythritol-2,4-cyclodiphosphate synthase: a target for antimicrobial therapy. Journal of medicinal chemistry 52: 2531-2542.

203. Rangarajan ES, Asinas A, Proteau A, Munger C, Baardsnes J, et al. (2008) Structure of [NiFe] hydrogenase maturation protein HypE from Escherichia coli and its interaction with HypF. Journal of bacteriology 190: 1447-1458.

204. Rao DR, Hariharan K, Vijayalakshmi KR (1969) A study of the metabolism of L-alpha gamma-diaminobutyric acid in a Xanthomonas species. The Biochemical journal 114: 107-115.

205. Rashid N, Kanai T, Atomi H, Imanaka T (2004) Among multiple phosphomannomutase gene orthologues, only one gene encodes a protein with phosphoglucomutase and phosphomannomutase activities in Thermococcus kodakaraensis. Journal of bacteriology 186: 6070-6076.

206. Robson RL, Morris JG (1974) Mobilization of granulose in Clostridium pasteurianum. Purification and properties of granulose phosphorylase. The Biochemical journal 144: 513-517.

207. Sabater B, Sebastian J, Asensio C (1972) Identification and properties of an inducible and highly specific fructokinase from Streptomyces violaceoruber. Biochimica et biophysica acta 284: 414-420.

208. Sagami H, Ogura K (1981) Geranylgeranyl pyrophosphate synthetase lacking geranyl-transferring activity from Micrococcus luteus. Journal of biochemistry 89: 1573-1580.

209. Sagami H, Ogura K (1985) Geranylpyrophosphate synthetase-geranylgeranylpyrophosphate synthetase from Micrococcus luteus. Methods in enzymology 110: 188-192.

210. Saha R, Dasgupta S, Basu G, Roy S (2009) A chimaeric glutamyl:glutaminyl-tRNA synthetase: implications for evolution. The Biochemical journal 417: 449-455.

211. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in Escherichia coli: genomic analyses and predictions. Proceedings of the National Academy of Sciences of the United States of America 97: 6652-6657.

212. Sancar A, Smith FW, Sancar GB (1984) Purification of Escherichia coli DNA photolyase. The Journal of biological chemistry 259: 6028-6032.

213. Sanchez M, Fernandez J, Martin M, Gibello A, Garrido-Pertierra A (1989) Purification and properties of two succinic semialdehyde dehydrogenases from Klebsiella pneumoniae. Biochimica et biophysica acta 990: 225-231.

214. Satoh S, Moritani C, Ohhashi T, Konishi K, Ikeda M (1994) Chloroplast ATPase in Acetabularia acetabulum: purification and characterization of chloroplast F1-ATPase. Bioscience, biotechnology, and biochemistry 58: 521-525.

215. Sazanov LA, Carroll J, Holt P, Toime L, Fearnley IM (2003) A role for native lipids in the stabilization and two-dimensional crystallization of the Escherichia coli NADH-ubiquinone oxidoreductase (complex I). The Journal of biological chemistry 278: 19483-19491.

216. Schauer S, Chaturvedi S, Randau L, Moser J, Kitabatake M, et al. (2002) Escherichia coli glutamyl-tRNA reductase. Trapping the thioester intermediate. The Journal of biological chemistry 277: 48657-48663.

217. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, et al. (2011) BRENDA, the enzyme information system in 2011. Nucleic acids research 39: D670-676.

218. Schmidt A, Wachtler B, Temp U, Krekling T, Seguin A, et al. (2010) A bifunctional geranyl and geranylgeranyl diphosphate synthase is involved in terpene oleoresin formation in Picea abies. Plant physiology 152: 639-655.

219. Schmidt LS, Sojka GA (1973) Enzymes of serine biosynthesis in Rhodopseudomonas capsulata. Archives of biochemistry and biophysics 159: 475-482.

220. Schryvers A, Lohmeier E, Weiner JH (1978) Chemical and functional properties of the native and reconstituted forms of the membrane-bound, aerobic glycerol-3-phosphate dehydrogenase of Escherichia coli. The Journal of biological chemistry 253: 783-788.

221. Schryvers A, Weiner JH (1981) The anaerobic sn-glycerol-3-phosphate dehydrogenase of Escherichia coli. Purification and characterization. The Journal of biological chemistry 256: 9959-9965.

222. Setlow P (1974) DNA polymerase I from Escherichia coli. Methods in enzymology 29: 3-12.

223. Seto D, Bhatnagar SK, Bessman MJ (1988) The purification and properties of deoxyguanosine triphosphate triphosphohydrolase from Escherichia coli. The Journal of biological chemistry 263: 1494-1499.

224. Sharma K, Chandra H, Gupta PK, Pathak M, Narayan A, et al. (2004) PknH, a transmembrane Hank's type serine/threonine kinase from Mycobacterium tuberculosis is differentially expressed under stress conditions. FEMS microbiology letters 233: 107-113.

225. Shepherd M, Dailey TA, Dailey HA (2006) A new class of [2Fe-2S]-cluster-containing protoporphyrin (IX) ferrochelatases. The Biochemical journal 397: 47-52.

226. Shin Y, Sawada K, Nagakura T, Miyanaga M, Moritani C, et al. (1996) Reconstitution of the F1-ATPase activity from purified alpha, beta, gamma and delta or epsilon subunits with glutathione S-transferase fused at their amino termini. Biochimica et biophysica acta 1273: 62-70.

227. Shinagawa E, Ano Y, Yakushi T, Adachi O, Matsushita K (2009) Solubilization, purification, and properties of membrane-bound D-glucono-delta-lactone hydrolase from Gluconobacter oxydans. Bioscience, biotechnology, and biochemistry 73: 241-244.

228. Shrivastava R, Ghosh AK, Das AK (2007) Probing the nucleotide binding and phosphorylation by the histidine kinase of a novel three-protein two-component system from Mycobacterium tuberculosis. FEBS letters 581: 1903-1909.

229. Silber KR, Keiler KC, Sauer RT (1992) Tsp: a tail-specific protease that selectively degrades proteins with nonpolar C termini. Proceedings of the National Academy of Sciences of the United States of America 89: 295-299.

230. Silva RG, de Carvalho LP, Blanchard JS, Santos DS, Basso LA (2006) Mycobacterium tuberculosis beta-ketoacyl-acyl carrier protein (ACP) reductase: kinetic and chemical mechanisms. Biochemistry 45: 13064-13073.

231. Silva RG, Rosado LA, Santos DS, Basso LA (2008) Mycobacterium tuberculosis beta-ketoacyl-ACP reductase: alpha-secondary kinetic isotope effects and kinetic and equilibrium mechanisms of substrate binding. Archives of biochemistry and biophysics 471: 1-10.

232. Simuth J, Zelinka J, Polek B (1975) Polynucleotide phosphorylase from Streptomyces aureofaciens: purification and properties. Biochimica et biophysica acta 379: 397-407.

233. Singh B, Lee CB, Sohng JK (2010) Precursor for biosynthesis of sugar moiety of doxorubicin depends on rhamnose biosynthetic pathway in Streptomyces peucetius ATCC 27952. Applied microbiology and biotechnology 85: 1565-1574.

234. Singh H (2010) Two decades with dimorphic Chloride Intracellular Channels (CLICs). FEBS letters 584: 2112-2121.

235. Sitthisak S, Knutsson L, Webb JW, Jayaswal RK (2007) Molecular characterization of the copper transport system in Staphylococcus aureus. Microbiology 153: 4274-4283.

236. Sivaraman H, Sivaraman C (1979) Cooperative binding of manganese to citrate lyase from Klebsiella aerogenes. FEBS letters 105: 267-270.

237. Sousa MC, McKay DB (2001) Structure of the universal stress protein of Haemophilus influenzae. Structure 9: 1135-1141.

238. Staab CA, Hellgren M, Grafstrom RC, Hoog JO (2009) Medium-chain fatty acids and glutathione derivatives as inhibitors of S-nitrosoglutathione reduction mediated by alcohol dehydrogenase 3. Chemico-biological interactions 180: 113-118.

239. Stadtman TC (1989) Clostridial glycine reductase: protein C, the acetyl group acceptor, catalyzes the arsenate-dependent decomposition of acetyl phosphate. Proceedings of the National Academy of Sciences of the United States of America 86: 7853-7856.

240. Steinbacher S, Kaiser J, Wungsintaweekul J, Hecht S, Eisenreich W, et al. (2002) Structure of 2C-methyl-d-erythritol-2,4-cyclodiphosphate synthase involved in mevalonate-independent biosynthesis of isoprenoids. Journal of molecular biology 316: 79-88.

241. Stricker O, Masepohl B, Klipp W, Bohme H (1997) Identification and characterization of the nifV-nifZ-nifT gene region from the filamentous cyanobacterium Anabaena sp. strain PCC 7120. Journal of bacteriology 179: 2930-2937.

242. Sudom A, Walters R, Pastushok L, Goldie D, Prasad L, et al. (2003) Mechanisms of activation of phosphoenolpyruvate carboxykinase from Escherichia coli by Ca2+ and of desensitization by trypsin. Journal of bacteriology 185: 4233-4242.

243. Suelter CH, Snell EE (1977) Monovalent cation activation of tryptophanase. The Journal of biological chemistry 252: 1852-1857.

244. Sukdeo N, Honek JF (2008) Microbial glyoxalase enzymes: metalloenzymes controlling cellular levels of methylglyoxal. Drug metabolism and drug interactions 23: 29-50.

245. Suzuki T, Zhang YW, Koyama T, Sasaki DY, Kurihara K (2006) Direct observation of substrate-enzyme complexation by surface forces measurement. Journal of the American Chemical Society 128: 15209-15214.

246. Switzer RL (1969) Regulation and mechanism of phosphoribosylpyrophosphate synthetase. I. Purification and properties of the enzyme from Salmonella typhimurium. The Journal of biological chemistry 244: 2854-2863.

247. Takahagi M, Iwasaki H, Shinagawa H (1994) Structural requirements of substrate DNA for binding to and cleavage by RuvC, a Holliday junction resolvase. The Journal of biological chemistry 269: 15132-15139.

248. Tang Y, Guest JR, Artymiuk PJ, Green J (2005) Switching aconitase B between catalytic and regulatory modes involves iron-dependent dimer formation. Molecular microbiology 56: 1149-1158.

249. Theze J, Kleidman L, St Girons I (1974) Homoserine kinase from Escherichia coli K-12: properties, inhibition by L-threonine, and regulation of biosynthesis. Journal of bacteriology 118: 577-581.

250. Thompson D, Simonson T (2006) Molecular dynamics simulations show that bound Mg2+ contributes to amino acid and aminoacyl adenylate binding specificity in aspartyl-tRNA synthetase through long range electrostatic interactions. The Journal of biological chemistry 281: 23792-23803.

251. Tisa LS, Sekelsky JJ, Adler J (2000) Effects of organic antagonists of Ca(2+), Na(+), and K(+) on chemotaxis and motility of escherichia coli. Journal of bacteriology 182: 4856-4861.

252. Torrecilla I, Leganes F, Bonilla I, Fernandez-Pinas F (2000) Use of recombinant aequorin to study calcium homeostasis and monitor calcium transients in response to heat and cold shock in cyanobacteria. Plant physiology 123: 161-176.

253. Traverso ME, Subramanian P, Davydov R, Hoffman BM, Stemmler TL, et al. (2010) Identification of a hemerythrin-like domain in a P1B-type transport ATPase. Biochemistry 49: 7060-7068.

254. Trujillo M, Ferrer-Sueta G, Radi R (2008) Kinetic studies on peroxynitrite reduction by peroxiredoxins. Methods in enzymology 441: 173-196.

255. Tsesin N, Kogan A, Gdalevsky GY, Himanen JP, Cohen-Luria R, et al. (2007) The structure of apo tryptophanase from Escherichia coli reveals a wide-open conformation. Acta crystallographica Section D, Biological crystallography 63: 969-974.

256. Tsuchiya D, Shimizu N, Tomita M (2008) Versatile architecture of a bacterial aconitase B and its catalytic performance in the sequential reaction coupled with isocitrate dehydrogenase. Biochimica et biophysica acta 1784: 1847-1856.

257. Tuske S, Singh K, Kaushik N, Modak MJ (2000) The J-helix of Escherichia coli DNA polymerase I (Klenow fragment) regulates polymerase and 3'- 5'- exonuclease functions. The Journal of biological chemistry 275: 23759-23768.

258. Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. Journal of molecular biology 283: 489-506.

259. Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17: 849-850.

260. van der Kaaij H, Desiere F, Mollet B, Germond JE (2004) L-alanine auxotrophy of Lactobacillus johnsonii as demonstrated by physiological, genomic, and gene complementation approaches. Applied and environmental microbiology 70: 1869-1873.

261. Verma JN, Goldfine H (1985) Phosphatidylserine decarboxylase from Clostridium butyricum. Journal of lipid research 26: 610-616.

262. Vertiev YV, Ezepchuk YV (1981) Purification and characterization of some enzymatic properties of neuraminidase from Corynebacterium ulcerans. Hoppe-Seyler's Zeitschrift fur physiologische Chemie 362: 1339-1344.

263. Villafane AA, Voskoboynik Y, Cuebas M, Ruhl I, Bini E (2009) Response to excess copper in the hyperthermophile Sulfolobus solfataricus strain 98/2. Biochemical and biophysical research communications 385: 67-71.

264. Wallon G, Yamamoto K, Kirino H, Yamagishi A, Lovett ST, et al. (1997) Purification, catalytic properties and thermostability of 3-isopropylmalate dehydrogenase from Escherichia coli. Biochimica et biophysica acta 1337: 105-112.

265. Walsh JP, Bell RM (1992) Diacylglycerol kinase from Escherichia coli. Methods in enzymology 209: 153-162.

266. Walters EM, Garcia-Serres R, Jameson GN, Glauser DA, Bourquin F, et al. (2005) Spectroscopic characterization of site-specific [Fe(4)S(4)] cluster chemistry in ferredoxin:thioredoxin reductase: implications for the catalytic mechanism. Journal of the American Chemical Society 127: 9612-9624.

267. Wandinger-Ness AU, Ness SA, Weiss RL (1986) Simultaneous purification of three mitochondrial enzymes. Acetylglutamate kinase, acetylglutamyl-phosphate reductase and carbamoyl-phosphate synthetase from Neurospora crassa. The Journal of biological chemistry 261: 4820-4827.

268. Wang SL, Fan KQ, Yang X, Lin ZX, Xu XP, et al. (2008) CabC, an EF-hand calcium-binding protein, is involved in Ca2+-mediated regulation of spore germination and aerial hypha formation in Streptomyces coelicolor. Journal of bacteriology 190: 4061-4068.

269. Wedler FC, Shreve DS, Kenney RM, Ashour AE, Carfi J, et al. (1980) Two glutamine synthetases from Bacillus caldolyticus, an extreme thermophile. Isolation, physicochemical and kinetic properties. The Journal of biological chemistry 255: 9507-9516.

270. Wiktorowicz JE, Bonner J (1982) Studies on histone acetyltransferase. Partial purification and basic properties. The Journal of biological chemistry 257: 12893-12900.

271. Wolff M, Seemann M, Tse Sum Bui B, Frapart Y, Tritsch D, et al. (2003) Isoprenoid biosynthesis via the methylerythritol phosphate pathway: the (E)-4-

hydroxy-3-methylbut-2-enyl diphosphate reductase (LytB/IspH) from Escherichia coli is a [4Fe-4S] protein. FEBS letters 541: 115-120.

272. Wu S, Shen R, Zhang X, Wang Q (2010) Molecular cloning and characterization of maltooligosyltrehalose synthase gene from Nostoc flagelliforme. Journal of microbiology and biotechnology 20: 579-586.

273. Yamasaki H, Moriyama T (1971) Delta-aminolevulinic acid dehydratase of Mycobacterium phlei. Biochimica et biophysica acta 227: 698-705.

274. Yang JK, Epstein W (1983) Purification and characterization of adenylate cyclase from Escherichia coli K12. The Journal of biological chemistry 258: 3750-3758.

275. Yang Y, Zhang M, Zhang H, Lei J, Jin R, et al. (2006) Purification and characterization of Mycobacterium tuberculosis indole-3-glycerol phosphate synthase. Biochemistry Biokhimiia 71 Suppl 1: S38-43.

276. Yao YF, Weng YM, Hu HY, Ku KL, Lin LL (2006) Expression optimization and biochemical characterization of a recombinant gamma-glutamyltranspeptidase from Escherichia coli novablue. The protein journal 25: 431-441.

277. Yen MR, Choi J, Saier MH, Jr. (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. Journal of molecular microbiology and biotechnology 17: 163-176.

278. Yen NT, Bogdanovic X, Palm GJ, Kuhl O, Hinrichs W (2010) Structure of the Ni(II) complex of Escherichia coli peptide deformylase and suggestions on deformylase activities depending on different metal(II) centres. Journal of biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry 15: 195-201.

279. Yeom SJ, Ji JH, Kim NH, Park CS, Oh DK (2009) Substrate specificity of a mannose-6-phosphate isomerase from Bacillus subtilis and its application in the production of L-ribose. Applied and environmental microbiology 75: 4705-4710.

280. Yoon SA, Ryu SI, Lee SB, Moon TW (2008) Purification and characterization of branching specificity of a novel extracellular amylolytic enzyme from marine

hyperthermophilic Rhodothermus marinus. Journal of microbiology and biotechnology 18: 457-464.

281. Zalkin H, Kling D (1968) Anthranilate synthetase. Purification and properties of component I from Salmonella typhimurium. Biochemistry 7: 3566-3573.

282. Zerez CR, Moul DE, Andreoli AJ (1986) NAD kinase from Bacillus licheniformis: inhibition by NADP and other properties. Archives of microbiology 144: 313-316.

283. Zhai Y, Saier MH, Jr. (2001) A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. Journal of molecular microbiology and biotechnology 3: 501-502.

284. Zheng R, Blanchard JS (2001) Steady-state and pre-steady-state kinetic analysis of Mycobacterium tuberculosis pantothenate synthetase. Biochemistry 40: 12904-12912.