**Gene regulatory network development for Pseudomonas Putida**

# By David Teixeira

# Registration: 201913863

# Supervisor: Dr Leighton Pritchard

# A thesis submitted in partial fulfilment for the degree of BSc (Hons) in Biochemistry

# This report is entirely my own work. Any information taken from others has been declared and referenced in the text

Intro: 843 words

Methods: 685 words

Results & Discussion

Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow G4 0RE, January 2023.

# **Technical Abstract**

Background: *Pseudomonas Putida* is a highly diverse bacterial species resulting from its pangenomes large accessory genome providing environmental adaptations. This diversity may cause unexpected issues when using *Pseudomonas Putida* in laboratory or industrial settings. Gene Regulatory Networks(GRNs) visualise how genes affect each other’s expression, producing GRNs *Pseudomonas Putida* strains provide models of what to expect when applying specific strains in laboratory or industry, optimising use.

Methods: Analyse nucleotide sequences of intergenic regions(IGRs) per genome, identify recurring elements indicative of specific function(e.g transcription factor(TF) binding site) using MEMEsuite software “MEME” and “MAST”, merge IGRs via self-made program on basis of being functionally similar on GRN and assume known properties of constituent IGRs shared by all in merged IGR.

Results: Network IGR node counts reduced 9-14%, Cases occurred where merged IGR contained IGR of known TF-binding profile allowing property expansion to all constituent IGRs further interconnecting the network. Networks remain discontinuous but to lesser degree.

Conclusions: To proceed 2000-5000 unidentified genes per genome need identification, alternative forms of gene expression regulation must be explored, and further TF-binding site data needed for *Pseudomonas Putida*.

# **Lay Abstract**

# **Intro**

Pseudomonas Putida is a highly diverse gram-negative rod-shaped species of the Pseudomonas genus[1]. Over 100 different strains have been identified in various environments throughout the world, resulting from a large ‘accessory genome’ providing different strains the necessary genes for survival in its environment[2, 3]. Its high adaptability to environments means it’s fairly easy to culture *in vitro* and its intrinsic metabolism is highly versatile with Pseudomonas Putida producing a range of cofactors needed for various reactions, making it optimal for use in laboratories and industrial manufacture of natural products and proteins[1, 4]. Beyond this, Pseudomonas Putida is an opportunistic pathogen known to infect the immunocompromised which can lead to bacteraemia and sepsis, albeit uncommonly[5, 6].

Cell activities are dictated, in part, by genes and the degree to which they’re expressed which is constantly subject to change. Likewise, gene expression is influenced by the products of genes – either directly by transcription factors(**TFs**), or indirectly by regulators of TFs[7, 8]. A gene regulatory network(**GRN**) is a way of visualising or modelling the way genes regulate each other for a species/genome in graph format wherein genes are represented by Nodes and directed edges are formed between genes if one gene has an effect on expression of the other, with the edge value representing if the effect is positive or negative[9, 10]. Alternatively, since bacterial genes commonly occur in Operons, bacterial GRN nodes may be Operons and the promoter controlling them with edges formed from promoter to operon(s) it controls and from operon to promoter where operon encodes a TF or regulator thereof associated with the promoter.

GRNs are extremely powerful tools developed in recent years thanks to advancements in computing and high-throughput screening[11]. They have many uses, the most obvious being a ‘genetic framework’ for cells – be this for designing accurate *in silico* cell models, or predicting the knock-on effect changing cell culture conditions *in vitro* will have on gene expression[11]. Theoretically GRNs could be used in industrial manufacture to maintain optimal gene expression patterns for production. For species with pangenomes, GRNs could be compared between different strains of the species and evolutionary relationship could be determined based on the similarities in promoter to gene/operon associations[11] – strains could be very similar genetically but differ phenotypically due to gene control by different promoters. GRNs can serve as the basis of various studies, be they pharmaceutical or general research, as exemplified by a 2019 paper by L.Lou *et al* wherein a partial regulatory network was created for Salmonella for the SPI-1 gene cluster which plays major roles in pathogenicity – the GRN may serve to identify drug targets for salmonellosis[12].

Bioinformatics is an ever expanding field combining computer science and biology wherein software is utilised to analyse biological information such as DNA sequences. There’s a huge range of applications for bioinformatic tools and, likewise, there is a huge range of tools publicly available for various purposes, however these are recent developments as prior technology wasn’t powerful enough for such tasks. This study utilises both self-made and publicly sourced bioinformatic tools for processing data from 222 NCBI sourced Pseudomonas Putida genomes:

* PROKKA is a Unix based tool used to annotate prokaryotic genomes, identifying all non-coding and coding regions, their locations and their predicted products[13].
* Roary is a Unix based tool that takes a collection GFF3 format annotated genomes(e.g PROKKA output) from individuals of the same species, determines which coding regions encode ‘full proteins’, and from those identifies which are core genes or accessory genes of the pangenome whilst recording relative location of genes in each genome they occur[14].
* PIGGY is a Unix based tool that emulates Roary but identifies the core and accessory *intergenic regions*(IGRs) from GFF3 format genomes, recording the relative location of IGRs in the genomes they occur[15].
* nhmmer is a Linux tool used to scan genomic data for specific nucleotide sequences, such as transcription factor(**TF**) binding sites from PIGGY identified IGRs[16].

These tools enable construction of an incomplete GRN – taking the identified gene and IGR outputs and using CollecTF TF binding data[17], IGRs that are TF binding sites can be identified; Python scripts can be designed to process these outputs, identify which genes encode TFs that binding sites were found for, and form directed graphs linking IGRs to genes/operons and genes/operons to TF-binding IGRs. 222 Pseudomonas Putida genomes underwent this process in a prior internship but produced incomplete GRNs due to lack of TF-binding site data and lack of accounting for indirect gene regulation via TF regulators.

Given the many uses a GRN for Pseudomonas Putida could have in both Industrial and laboratory settings, as well as the fact GRNs are an emerging tool only made possible in recent years thanks to bioinformatics developments, this project aims to build upon the incomplete networks produced from 222 Pseudomonas Putida genomes by identifying TF binding sites missed previously and potentially incorporating regulators of TFs that indirectly influence gene expression to produce accurate GRNs for various Putida strains. These GRNs could serve as foundations for realistic Putida cell modelling and provide valuable information for optimising industrial use of Putida strains.

# **Methods**

Given the computational nature of the project, much of the processes involve self-made or sourced software. At the start of the project, the state of the GRNs consisted of IGRs linked to the operon(s) they regulate, and operons to IGR(s) they produce a transcription factor for according to collecTF data.

# ***Motif analysis***

A lack of further transcription factor(TF) binding data limits network development by forming links from *operon to IGR*, so the initial focus was to reduce graph clutter. IGRs were identified in the genomes using PIGGY which returned *whole* regions between genes and grouped them by sequence similarity under ID numbers, whilst TFs only bind specific motifs within IGRs with some degree of sequence variation[18] – therefore IGRs with similar TF-binding profiles may fall under separate “IDs” due to variations in the non-TF binding sections of the IGR.

By identifying potential TF binding motifs in IGR sequences, we may group different IGR IDs by similar TF-binding profiles according to similar motif composition, this may lead to IGRs being re-classified as TF binding sites for a TF we have binding data for, or may group IGRs that bind the same *unknown* TF. MEMEsuite is a collection of motif analysis tools, allowing motif discovery from sequences and location of specific motifs in sequences[19]:

1. For each genome(n=222) a multifasta was produced using self-made python script to collate all IGR sequences in that genome from the collated IGR\_sequences.txt produced by PIGGY. Following this each multifasta was run through MEME, each run searching for 50 non-overlapping motifs occurring any number of times per sequence.
2. MAST used the resulting MEME.HTML file for each genome to identify occurrences of each motif in the IGR multifasta corresponding to that genome, producing a MAST.txt file for each genome that lists IGRs bearing atleast one of the motifs alongside motif composition of said IGRs.
3. The MAST.txt and original GRN were fed into a self-made python script *Motif\_comparator\_V2.4.py* which compared the motif composition of all IGRs for a genome for similarity using a scoring system accounting for number of shared motifs, difference in number of motifs & interspersing-regions, relative motif order, and number of shared interspersing regions and their order. IGRs deemed similar were then merged in the original GRN to form a single node with all edges of the previous constituent IGRs.

# ***Regulon variation analysis***

A side-project to analyse IGR to gene association variation across genomes was conducted given network development cannot progress much further without TF binding data.

Due to their respective IGRs bearing the same TF-binding sites, multiple genes may be under regulation by the same transcription factor; For each TF, the collection of genes regulated is called a regulon and typically the constituent genes are related functionally[20, 21].

Between two otherwise genetically similar strains, differing regulon composition across strains may be responsible for phenotypic differences like environmental adaptations/responses to environments under identical conditions.

To analyse composition variation, 2 self-made python script were developed:

1. *Locustag\_genename\_IGR\_associator.py* which reads PROKKA .gbk files, operon predictions from internship data and IGR-to-Gene association data from PIGGY. Outputs “Genelist” file per genome, listing every gene in the genome alongside operon it belongs to, name (if known) and associated IGR.
2. *Cluster-gene\_association\_comparator\_V4.py* which reads aforementioned “Genelists”, nhmmer outputs for TF-binding sites and the PIGGY IGR\_presence\_absence.csv to create the various files, with statistics where appropriate (%Genes associated to “x” IGR, IGR to %operons with “x” gene, etc). Only relevant files listed below:
   1. *Collated\_gene\_to\_igr\_associations.txt* Listing all genes and the typical IGR associations across all genomes.
   2. *Collated\_bsigr\_to\_gene\_associations.txt* Listing each IGR for which TF-binding profile is known and the typical associated operon composition across all genomes.
   3. *Gene\_to\_bsigr\_variation\_analysis.txt* Listing each *named* gene that associates with >1 TF-binding IGR across all genomes for which the TF binding profiles differ.

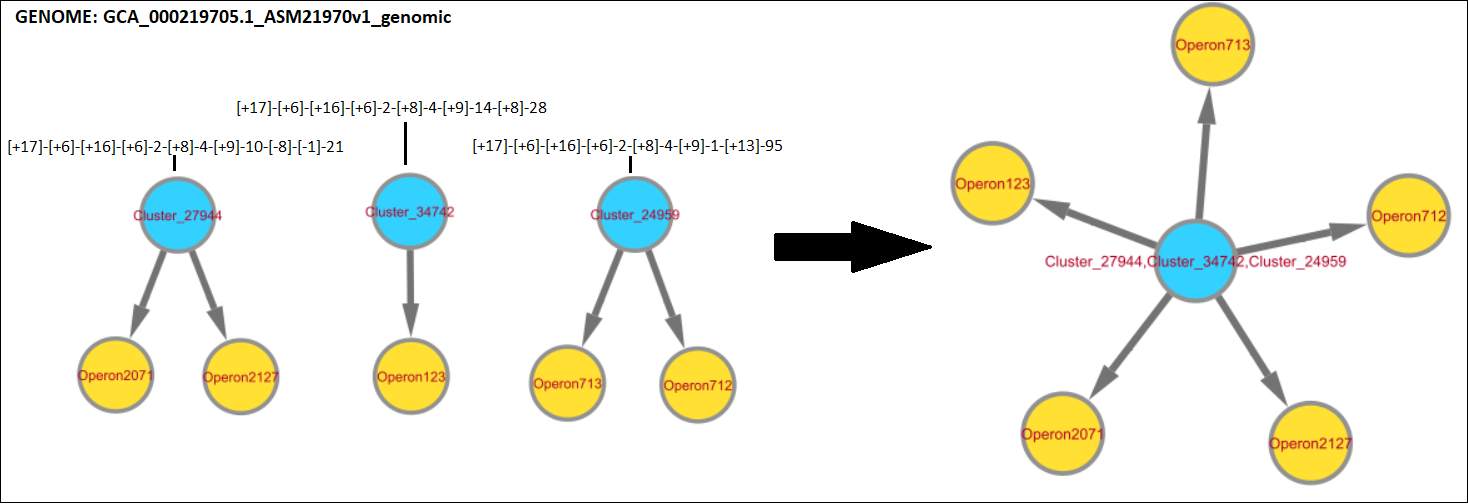
Taking the gene\_to\_bsigr\_variation\_analysis.txt, the regulatory mechanisms (Positive or negative regulation, mechanism(s) of activation) of all TFs listed were manually reviewed alongside the function of the named genes found regulated by those TFs, hypothesis were made as to why gene to TF association(regulon) may vary across the genomes and what this suggests about the phenotype and environment of the strain.

# **Results and Discussion**

# ***Motif analysis***



*Figure 1 – Edited views of MEME & MAST outputs using IGR multifasta extracted for given genome from PIGGY output(Genbank accession GCA\_000219705.1)* ***A.*** *MEME.HTML partial view depicting 17 of 50 discovered motifs with e-value < 0.05, 18th motif onwards greyed out due to e-value > 0.05.* ***B.*** *MAST.txt partial view depicting the motif composition of some IGRs when searching for motifs identified by MEME in the IGR multifasta.* ***C.*** *Motifs searched for in IGR multifasta, for motifs with >60% similarity only the one with lowest e-value was searched for, the rest omitted.*

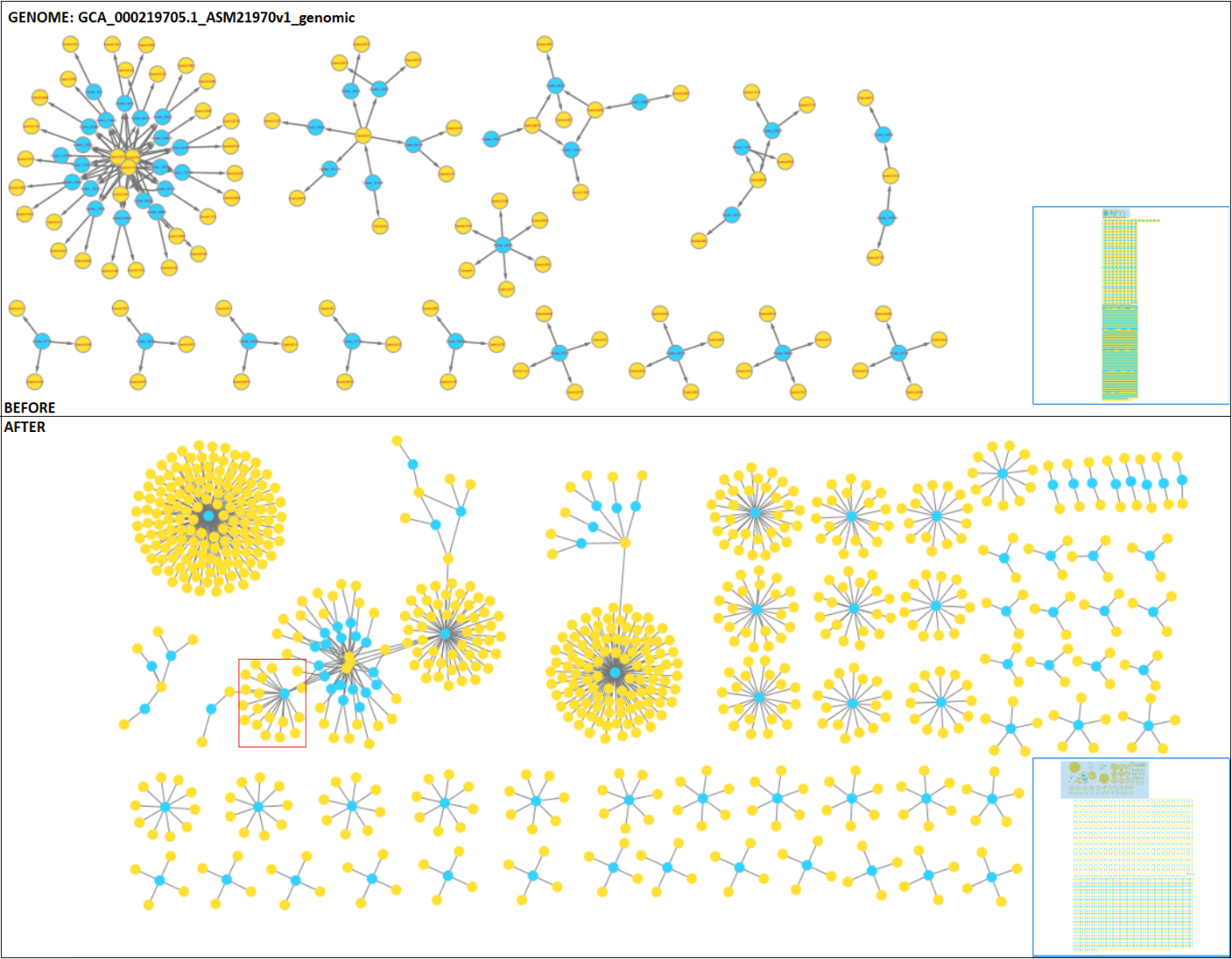


*Figure 2 – Edited Cytoscape view of a section of graph produced from sequence data (Genbank accession GCA\_000219705.1) before and after motif analysis. Blue nodes represent IGRs, yellow nodes represent operons, arrows represent direction of relationship(IGR regulates operon/Operon product binds IGR). Strings above IGR nodes represents motif composition (of one occurrence of IGR, if multiple occurrences applicable) according to MAST using MEME output as query, numbers in [] correspond to a specific motif, numbers outside [] represent the number of nucleotides between motifs, +/- represent strand.*

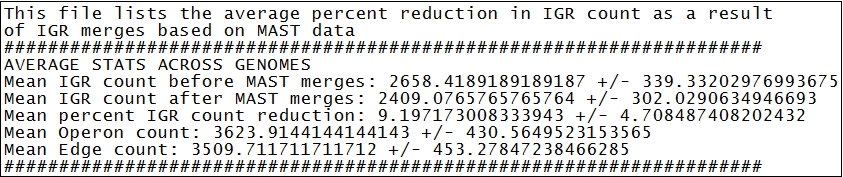
As shown in figure 1, MEME successfully identified motifs from IGR sequences that could then be found again in the same IGR sequences using MAST. On average, 15-16 motifs per genome were identified with e-values < 0.05, of which between 1-2 were omitted due to similarity to other motifs. These motifs could then be applied to MAST to identify motif occurrences within the same IGRs from which they were discovered, seemingly successfully as for IGRs with multiple occurrences within the same genome the motif composition of each occurrence was near identical with minor variation in interspersing regions.

Figures 2 depicts the result of merging IGRs based on motif composition. In figure 2, Cluster\_27944 is the only IGR with multiple occurrences within GCA\_000219705.1 with both occurrences having the same motif structure further consolidating the fact functionally similar IGRs have similar motif composition. Therefore, the fact that Clusters 34742 and 24959 only differ towards the right of motif [+9] suggests similar functionality. It is *possible* the variation following motif [+9] corresponds to different regulator binding sites allowing selective regulation of operons downstream of each IGR despite expression being determined by TF binding to the conserved structure preceding [+9].

However this raises a possible issue; This assumes that if most motifs shared between IGRs are identical that the same TF binds, however it may be the case that a TF only binds 1/2 motifs out of the overall IGR and these motifs may not be present in all merged IGRs, thus assuming TF-binding profiles are shared is invalid. It is possible that the motifs following [+9] are such TF binding motifs, with the motifs prior to [+9] being *general* binding sites *necessary* for protein binding in transcription.



*Figure 3 – Cytoscape view of regulatory network produced from sequence data (Genbank accession GCA\_000219705.1) before and after IGR merging according to motif comparison. Blue nodes represent IGRs, yellow nodes represent operons. Due to size constraints, only sections of the network with a merged IGR node and/or >3 interconnected nodes are depicted. 222 networks produced total, 20-150 IGRs merged on average per network. The blue box is a map of the network, the blue rectangle representing the current area shown. Red box contains section wherein merged node contains IGR of known TF-binding profile.*



*Figure 4 – Partial view of V2\_GRN\_both\_stats.txt showing average IGR/Operon/Edge counts +/- standard deviation across all genomes, as well as average percent reduction to IGR node count as a result of merges based on MAST results. Percent reduction represents disparity between IGR node count before and after MAST merges measured as a percentage of the IGR node count before.*

Figure 3 depicts the same as figure 2 but on a larger scale. A significant reduction to node count was observed in each GRN, averaging to 9.2 +/- 4.7 % as seen in figure 4. For most GRNs this caused the formation of 2 or more highly interconnected network segments as a result of one of the merged IGRs containing a TF binding site, thus that property being extended to all IGRs it was merged with. However generally most merges resulted in a single IGR node connected to many operons whereupon the network section terminates.

Whilst the reduction was fairly small given the scale of the graph, it’s to be expected given that PIGGY should’ve grouped most IGRs correctly beforehand. The merging should have a more significant effect in the future once further TF binding site data is obtained as if one IGR is a binding site for “X” transcription factor, it may be extrapolated that all IGRs merged with that one are aswell – This can already be seen in figure 3 after merging in the red box where the merged IGR node contains *Cluster\_11578*,a known LexA binding site, resulting in all other merged IGRs being assumed to be the same linking their operons to the LexA regulon.

*Issues with motif analysis results*

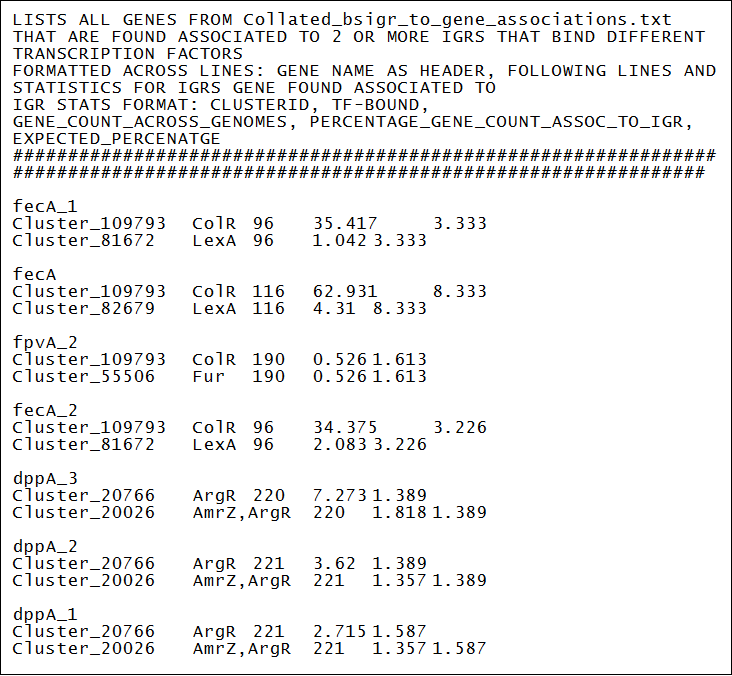
Whilst the merging process was successful, there is an issue with the networks themselves in that the networks fail to account for how genes could *indirectly* affect one another’s expression. Currently, the network only assumes that genes encoding transcription factors can affect another genes expression given that is the function of transcription factors, however a gene product that interferes with the function of transcription factors could be said to affect the expression of all genes that transcription factor regulates. As described in a 1996 paper by C.F.Calkhoven *et al*, transcription factors have various modes of regulation like ligand binding, phosphorylation or other post-translational modification[22] – any genes that are involved in phosphorylation, dephosphorylation, modification or ligand production for this transcription factor could be said to indirectly influence expression of the genes under that transcription factors regulon by positive or negative interference of transcription factor function.

For example, a gene may encode a ligand that initiates a signal transduction cascade that activates/inactivates “X” TF, thus indirectly regulates expression of “X” TF associated genes. Furthermore, the pathways that lead to transcription factor activation aren’t always linear. Whilst rare for bacteria as their signal transduction pathways for transcription factors are shorter than eukaryotes[23], in more complex signalling pathways an upstream element of the pathway may have multiple targets that ultimately lead to activation and/or inactivation of multiple different transcription factors. For example in eukaryotes, activation of the MAPK cascade leading to JNK activation can also lead to ERK1/2 phosphorylation, it will be difficult to account for all such occurrences and complicate the networks significantly[24].

Additionally, this network operates on the basis that gene expression is controlled solely through transcription, however this isn’t the case. Gene expression is defined as the process wherein genetic information is used to drive production of a molecule like protein or RNA[25], which encompasses both transcription, translation and possibly post-translational modification. Regulation mechanisms exist at each stage, for example siRNAs which regulate translation by binding complementary mRNA molecules leading to their destruction thus preventing the mRNA driving protein production[26], thus even if all TF’s and TF-regulating genes are accounted for, the network will still be incomplete. However such a detailed network may be to complicated for use, and for some uses just accounting for TF’s and TF regulators may be sufficient.

Finally, each genome had between 2000-5000 genes encoding unknown products. This may have resulted from an issue in PROKKA annotation and severely limits network development given that, in some networks, TF binding sites were identified yet the genome appeared not to encode the corresponding TF. Identifying these genes should be a priority should network development proceed, as it should yield immediate improvements if a previously unidentified TF was discovered, and the unknown genes may contain indirect regulators of gene expression like siRNA’s or TF regulators as previous discussed.

# ***Regulon variation analysis***



*Figure 5 – Partial view of Gene\_to\_bsigr\_variation\_analysis.txt produced by Cluster-gene\_association\_comparator\_V4.py, data entries organised vertically with each entry separated by whitespace line. Entries begin with gene name followed by each IGR for which TF binding profile is known and the gene is found regulated by, across all genomes. From left to right the format is IGR, TF bound, Total number of occurrences of gene, Percentage of total gene count found associated to IGR, Percentage of total gene count expected assuming equal distribution across all IGRs it’s found associated to.*

Figure 5 depicts part of the Gene\_to\_bsigr\_variation\_analysis.txt produced by self-made python script to collate variation to TF regulation of genes across genomes. This data was manually processed, researching regulatory mechanisms of the relevant TFs to draw hypothesis on what this means for phenotypic variation between genomes under identical conditions. Given the length of the file and the amount of research that must go into each TF and gene, so far only a few regulon variations have been studied – 3 interesting examples are shown below.

*fecA regulon variation*

First studied was *fecA*, which was found 116 times across all genomes, of which 63% were associated to Cluster\_109793 under the **ColR** regulon, 4.3% to Cluster­\_82679 under the **LexA** regulon and the remaining 32.7% under unknown regulons.

**ColR** is the effector of a two-component system activated by sensor ColS in response to Zinc, among other cations, and is an *activator* that upregulates genes involved in metal ion homeostasis and membrane functionality[27, 28]. **LexA** on the other hand is a constitutively activated *repressor* that downregulates DNA repair and error checking genes, only being deactivated by autocatalytic cleavage in response to DNA damage[29, 30].

Given that *fecA* is an Iron dicitrate transporter that, in *Pseudomonas Aeruginosa*, is upregulated in times of iron deficiency[31] it makes sense for it to fall under the **ColR** regulon given that Fe2+ activates **ColR** therefore, for genomes this regulon occurs, Iron can directly induce its own uptake, however there must be some negative feedback system to prevent overload of Iron.

In genomes where *fecA* falls under the **LexA** regulon, it’s possible the genome source is in a high iron environment where its more crucial to suppress than promote Iron uptake given the cell obtains enough via other mechanisms. However, given that Fe2+ catalyses ROS production[32], the fact that *fecA* falls under **LexA** regulation is confusing given that ROS may damage DNA which would deactivate **LexA**, leading to upregulation of *fecA* and Iron uptake, exacerbating the DNA damage – This may indicate an error, imply **LexA** regulation is more complex than previously believed, or more unlikely this regulon structure could serve as a rudimentary form of apoptosis, killing the cell via further Fe2+ uptake promoting ROS-induced cell damage to prevent the damaged DNA being transferred.

*czcR\_3 regulon variation*

After that was *czcR\_3* – the effector component of a two-component system wherein *czcS* detects high concentration of ions like Zn2+, Cu2+, Co2+ and Cd2+ in the periplasmic space and activates *czcR\_3*in response, allowing it to promote expression of metal ion efflux pump *czcCBA* which exports cations from the cytoplasm and periplasm to the extracellular space[33-35]. With 182 occurrences of *czcR\_3* across all genomes, 17% were associated to Cluster\_66703 under the **ColR** regulon and 0.55% were associated to Cluster\_68396 under the **CueR** regulon, the remaining ~82.5% under unknown regulons.

**ColR** regulation of *czcR\_3* makes sense given that **ColR** is activated by some of the same ions *czcCBA* exports – by upregulating *czcR\_3*, *czcR\_3* can promote expression of *czcCBA* to promote efflux of said ions in a negative-feedback-like manner, albeit a convoluted one, to maintain homeostasis[27, 28].

**CueR** is a transcriptional activator involved in lowering cytoplasmic Cu2+ concentration by promoting expression of Cu2+ sequestering proteins upon binding Cu2+ itself[36]. Again, given that Cu2+ both leads to expression of *czcR\_3* and is exported by *czcCBA*, it could serve as a homeostatic mechanism for cytoplasmic/periplasmic cation concentrations.

It is possible the reason for *czcR\_3* falling under different regulons is that in genomes where *czcR\_3* falls under the **ColR** regulon, environmental Cu2+ is ubiquitous thus the cell has no issues with Cu2+ deficiency and can export freely to avoid overload, whereas where *czcR\_3* falls under the **CueR** regulon, environmental Cu2+ may be scarce thus to avoid exporting what limited Cu2+ the cell has it will only express *czcR\_3* (and by proxy *czcCBA*) when Cu2+ alone is high in concentration. This theory is pure conjecture and poses many issues such as how the genome containing the *czcR\_3* **CueR** regulon occurs handles high concentrations of cations other than Cu2+ and the fact that Cu2+ can activate **CueR** even at lower concentrations thus **CueR** regulation of *czcR\_3* may not solve issues of Cu2+ scarcity.

*oprD\_6 regulon variation*

The gene *oprD\_6* was found 213 times across all genomes, with 7% associated to Cluster\_4421 under the **ArgR** regulon, 0.5% associated to Cluster\_11358 under the **ColR** regulon and 0.5% associated to Cluster\_97363 under the **Zur** regulon, with the remaining 92% falling under unknown regulons. *oprD\_6* encodes an outer membrane porin which functions as a channel for passive transport of basic amino acids and select small peptides, but it is also a means of entry for antibiotics like carbapenem[37].

**ArgR** is both a transcriptional activator and repressor – when activated by arginine binding in *Pseudomonas Aeruginosa*, it upregulates Arginine catabolism genes and represses Arginine synthesis genes, essentially acting to maintain Arginine homeostasis by reducing Arginine concentration in cytosol[38-41]. **ArgR** normally functions to reduce intracellular Arginine concentration, therefore since *oprD\_6* functions as a channel for basic amino acids like Arginine it’s likely **ArgR** acts as a repressor for *oprD\_6*, suppressing *oprD\_6* to prevent further Arginine uptake.

**Zur** is a transcriptional repressor of genes involving Zinc influx and is activated through binding Zinc, therefore functions as a homeostatic mechanism to lower Zinc concentration[42]. Similarly to **ArgR**, since *oprD\_6* functions as a basic amino acid channel and basic amino acids bear positive charge *oprD\_6* may also enable cation flux like Zn2+ given the charge similarity and small size, so *oprD\_6* may fall under the **Zur** regulon in genomes to act as a negative-feedback-like system for Zn2+ uptake through *oprD\_6* repression – however there doesn’t seem to be an obvious benefit for *oprD\_6* falling under the **Zur** regulon vs the **ArgR** regulon, unless the small size of Zn2+ and its additional routes of entry means it can achieve higher concentrations in shorter times than Arginine, therefore making **Zur** active more often than **ArgR** thus conferring slightly greater resistance to antibiotics like carbapenem, given that less *oprD\_6* means less routes of entry.

In genomes where *oprD\_6* falls under the **ColR** regulon, it’s possible the genome source hasn’t been subjected to antibiotics like carbapenem as **ColR** is a transcriptional *activator* activated by increased concentration of environmental Zn2+ and other cations – since *oprD\_6* enables antibiotic entry, the fact it isn’t under a repressors regulon in this strain suggests either that: 1). The strain hasn’t encountered such antibiotics thus doesn’t benefit from *oprD\_6* downregulation, 2). The strains other means of antibiotic defence are sufficient or 3). The **ColR** binding IGR Cluster\_11358 also binds a repressor not yet identified.

*Issues with Regulon variation analysis*

Whilst an interesting project, these findings currently serve little value other than a proof of concept for phenotypic variation analysis in genetically similar strains. When the networks are provided more TF binding data, this technique may serve more uses such as identifying aberrant strain responses to different factors when performing *in vitro* cultures or potentially in researching the phylogenetic relationship between strains where other methods prove insufficient.

Additionally, the analysis doesn’t yet account for IGR merging from motif analysis. In the script used, genes were only listed when the IGRs they’re associated across all genomes to had known TF binding profile *and* atleast two of said IGRs bound different TF’s – were I to account for merged IGRs, for each merged IGR the TF binding properties of one constituent IGR would be extended to all constituent IGRs thus potentially allowing more genes to meet the criteria described above for being included in the analysis.

# **Conclusion**

As a whole, the project was successful – IGR motif analysis and subsequent merges reduced IGR node count around 10%, in some cases identifying previously overlooked TF-binding sites by extending TF-binding properties of one constituent IGR of a merged IGR node to all constituents, and will help identify overlooked TF-binding IGRs in the future once more TF-binding data is obtained. Furthermore, whilst not yielding much data regulon variation analysis was also successful, though the findings serve only as a proof of concept for epigenetic and phylogenetic analyses.

This paper also identified some next steps that should be taken if one wants to expand upon the GRNs – mainly being accounting for regulators of TF’s when forming links in the network, accounting for modes of expression regulation outside transcription, identifying the remaining unidentified genes in each genome, validation to prove that merged IGRs constituents share TF binding profiles and simply obtaining more TF binding site data for Pseudomonas Putida.

# **References**

1. Poblete-Castro, I., et al., *Industrial biotechnology of Pseudomonas putida and related species.* Applied Microbiology and Biotechnology, 2012. **93**(6): p. 2279-2290.

2. Silby, M.W., et al., *Pseudomonas genomes: diverse and adaptable.* FEMS Microbiology Reviews, 2011. **35**(4): p. 652-680.

3. Udaondo, Z., et al., *Analysis of the core genome and pangenome of Pseudomonas putida.* Environ Microbiol, 2016. **18**(10): p. 3268-3283.

4. Martin-Pascual, M., et al., *A navigation guide of synthetic biology tools for Pseudomonas putida.* Biotechnology Advances, 2021. **49**: p. 107732.

5. Horii, T., H. Muramatsu, and Y. Iinuma, *Mechanisms of resistance to fluoroquinolones and carbapenems in Pseudomonas putida.* Journal of Antimicrobial Chemotherapy, 2005. **56**(4): p. 643-647.

6. Chamon, R.C., et al., *KPC-2 producing Pseudomonas putida as an unexpected pathogen of catheter-associated bloodstream infection.* J Infect Dev Ctries, 2020. **14**(4): p. 411-414.

7. Lambert, S.A., et al., *The Human Transcription Factors.* Cell, 2018. **172**(4): p. 650-665.

8. Mitsis, T., et al., *Transcription factors and evolution: An integral part of gene expression (Review).* World Acad Sci J, 2020. **2**(1): p. 3-8.

9. Aalto, A., et al., *Gene regulatory network inference from sparsely sampled noisy data.* Nature Communications, 2020. **11**(1): p. 3493.

10. Koutrouli, M., et al., *A Guide to Conquer the Biological Network Era Using Graph Theory.* Frontiers in Bioengineering and Biotechnology, 2020. **8**.

11. Emmert-Streib, F., M. Dehmer, and B. Haibe-Kains, *Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks.* Frontiers in Cell and Developmental Biology, 2014. **2**.

12. Lou, L., et al., *Salmonella Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network.* Front Cell Infect Microbiol, 2019. **9**: p. 270.

13. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-2069.

14. Page, A.J., et al., *Roary: rapid large-scale prokaryote pan genome analysis.* Bioinformatics, 2015. **31**(22): p. 3691-3693.

15. Thorpe, H.A., et al., *Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria.* GigaScience, 2018. **7**(4).

16. Wheeler, T.J. and S.R. Eddy, *nhmmer: DNA homology search with profile HMMs.* Bioinformatics, 2013. **29**(19): p. 2487-2489.

17. Kiliç, S., et al., *CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria.* Nucleic Acids Res, 2014. **42**(Database issue): p. D156-60.

18. Inukai, S., K.H. Kock, and M.L. Bulyk, *Transcription factor-DNA binding: beyond binding site motifs.* Curr Opin Genet Dev, 2017. **43**: p. 110-119.

19. Bailey, T.L., et al., *The MEME Suite.* Nucleic Acids Research, 2015. **43**(W1): p. W39-W49.

20. Lozada-Chávez, I., S.C. Janga, and J. Collado-Vides, *Bacterial regulatory networks are extremely flexible in evolution.* Nucleic Acids Res, 2006. **34**(12): p. 3434-45.

21. Erill, I., et al., *Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics.* Nucleic Acids Res, 2004. **32**(22): p. 6617-26.

22. Calkhoven, C.F. and G. Ab, *Multiple steps in the regulation of transcription-factor level and activity.* Biochem J, 1996. **317 ( Pt 2)**(Pt 2): p. 329-42.

23. Seshasayee, A.S.N., K. Sivaraman, and N.M. Luscombe, *An Overview of Prokaryotic Transcription Factors*, in *A Handbook of Transcription Factors*, T.R. Hughes, Editor. 2011, Springer Netherlands: Dordrecht. p. 7-23.

24. Zhang, W. and H.T. Liu, *MAPK signal pathways in the regulation of cell proliferation in mammalian cells.* Cell Research, 2002. **12**(1): p. 9-18.

25. Buccitelli, C. and M. Selbach, *mRNAs, proteins and the emerging principles of gene expression control.* Nature Reviews Genetics, 2020. **21**(10): p. 630-644.

26. Dana, H., et al., *Molecular Mechanisms and Biological Functions of siRNA.* Int J Biomed Sci, 2017. **13**(2): p. 48-57.

27. Ainsaar, K., et al., *The ColRS signal transduction system responds to the excess of external zinc, iron, manganese, and cadmium.* BMC Microbiology, 2014. **14**(1): p. 162.

28. Kivistik, P.A., et al., *The ColRS two-component system regulates membrane functions and protects Pseudomonas putida against phenol.* J Bacteriol, 2006. **188**(23): p. 8109-17.

29. Abella, M., et al., *Cohabitation of two different lexA regulons in Pseudomonas putida.* J Bacteriol, 2007. **189**(24): p. 8855-62.

30. Calero, S., X. Garriga, and J. Barbé, *Analysis of the DNA damage-mediated induction of Pseudomonas putida and Pseudomonas aeruginosa lexA genes.* FEMS Microbiol Lett, 1993. **110**(1): p. 65-70.

31. Marshall, B., et al., *Citrate-mediated iron uptake in Pseudomonas aeruginosa: involvement of the citrate-inducible FecA receptor and the FeoB ferrous iron transporter.* Microbiology, 2009. **155**(1): p. 305-315.

32. Pasqua, M., et al., *Ferric Uptake Regulator Fur Is Conditionally Essential in Pseudomonas aeruginosa.* J Bacteriol, 2017. **199**(22).

33. Liu, H., et al., *The Connection between Czc and Cad Systems Involved in Cadmium Resistance in Pseudomonas putida.* Int J Mol Sci, 2021. **22**(18).

34. Dieppois, G., et al., *The transcriptional regulator CzcR modulates antibiotic resistance and quorum sensing in Pseudomonas aeruginosa.* PLoS One, 2012. **7**(5): p. e38148.

35. Liu, P., et al., *The Role of CzcRS Two-Component Systems in the Heavy Metal Resistance of Pseudomonas putida X4.* Int J Mol Sci, 2015. **16**(8): p. 17005-17.

36. Novoa-Aponte, L., D. Ramírez, and J.M. Argüello, *The interplay of the metallosensor CueR with two distinct CopZ chaperones defines copper homeostasis in Pseudomonas aeruginosa.* J Biol Chem, 2019. **294**(13): p. 4934-4945.

37. Li, H., et al., *Structure and function of OprD protein in Pseudomonas aeruginosa: from antibiotic resistance to novel therapies.* Int J Med Microbiol, 2012. **302**(2): p. 63-8.

38. Lu, C.D., et al., *The ArgR regulatory protein, a helper to the anaerobic regulator ANR during transcriptional activation of the arcD promoter in Pseudomonas aeruginosa.* J Bacteriol, 1999. **181**(8): p. 2459-64.

39. Schaumburg, C.S. and M. Tan, *Arginine-dependent gene regulation via the ArgR repressor is species specific in chlamydia.* J Bacteriol, 2006. **188**(3): p. 919-27.

40. Charlier, D. and I. Bervoets, *Regulation of arginine biosynthesis, catabolism and transport in Escherichia coli.* Amino Acids, 2019. **51**(8): p. 1103-1127.

41. Lu, C.D., Z. Yang, and W. Li, *Transcriptome analysis of the ArgR regulon in Pseudomonas aeruginosa.* J Bacteriol, 2004. **186**(12): p. 3855-61.

42. Ducret, V., et al., *A Zur-mediated transcriptional regulation of the zinc export system in Pseudomonas aeruginosa.* BMC Microbiol, 2023. **23**(1): p. 6.