DEMYSTIFYING LANTHANOMES: A PANGENOMIC APPROACH

By

Rian Rafsan

21276004

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of

Masters Of Science in Biotechnology

School Of Data & Sciences

BRAC University

April 2023

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing a degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

_____

Rian Rafsan

21276004

Approval

The thesis/project titled ***"Demystifying Lanthanomes: A Pangenomic Approach"*** submitted by Rian Rafsan (21276004) of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Masters of Science in Biotechnology on 27th April, 2023.

Examining Committee:

Supervisor:
(Member)

Dr. Iftekhar Bin Nasser
Associate Professor, MNS
BRAC University

Co-Supervisor:
(Member)

Tushar Ahmed Shishir
Lecturer, MNS
BRAC University

Program Director:
(Member)

Dr. Munima Haque
Associate Professor, MNS
BRAC University

External Expert Examiner:
(Member)

Dr. Zahid Hayat Mahmud
Scientist & Head,
LSSD, icddr,b

Departmental Head:
(Chair)

Dr. Yusuf Haider
Professor, MNS
BRAC University

Ethics Statement

I hereby declare that I did not conduct any sort of unethical action whatsoever as a means to perform my thesis work. I remained wholeheartedly sincere, honest and worked with utmost dedication to achieve this work. No living organism was hurt or killed while conducting this research.

Rian Rafsan

Student ID: 21276004

Abstract

Certain members of multi-generic, gram negative proteobacteria (newly named *pseudomonatota*) have recently been reported to be capable of successfully utilizing the catalytic potential of lanthanide metals for the efficient dehydrogenation of alcohols as a part of their respective regular metabolism. While recent studies did essentially help the characterization of most genes responsible for such lanthanomic behavior, much of the knowledge space regarding lanthanomes is still engulfed in a void. With an aim to potentially reduce this knowledge gap we conducted an *in silico* pangenomic analysis on 31 non redundant strains spanning 23 genera, who were previously proven to be lanthanomes in a typical wet lab. We tried to confirm the lanthonomicity of these microorganisms by screening the resultant pangenome for lanthanide dependent genes- *xoxF*, *xoxG*, *ExaF*, *PedH*, *fldA* & *gfaA*. The analysis resulted in 181 core, 292 soft core, 7894 shell & 26805 cloud gene clusters respectively, while 26 of the 31 strains showed the presence of at least 1 of the selected lanthanomic genes. The results also support the heterogenic and subsequently divergent behavior of lanthanomes alongside events of concurrent clustering & genetic duplication.

A true chef is one whose dish is his own reflection,

One who pursues cuisine to the utmost perfection,

One who can answer the question with absolute honesty- ***What have you cooked?***

-Senzaemon Nakiri (Food Wars!:Shokugeki no Soma)

Acknowledgements

I would like to take this opportunity to first & foremost pay my utmost thanks and eternal gratitude to late Professor Dr. Ziauddin Ahmed, former departmental head & chairperson, department of Mathematics & Natural Sciences (MNS), BRACU for his constant encouragement and providing me with the opportunity to pursue my studies in biotechnology.

I would also like to express my cordial thanks to professor Dr. Mahbub Majumder; dean, School of Data & Sciences for his able help and guidance.I am greatly thankful to professor Dr. AFM Yusuf Haider; chairperson, department of Mathematics & Natural Sciences for his kind support and benevolent patronage throughout my journey as a student pursuing my masters degree.

I would especially want to thank Professor Dr. Aparna Islam for her constant support; both from academic and mental perspective who took me under her wing and made sure I didn't digrace from my path in the endeavor of knowledge.

I am deeply grateful to my supervisor Dr. Iftekhar Bin Nasser, associate professor, department of life sciences BRACU and my co-supervisor Mr. Tushar Ahmed Shishir, lecturer at department of life sciences BRACU respectively for their able guidance & constructive insights throughout my thesis journey ending in a successful completion.

I am also thankful to Mr. Rafeed Rahman Turjya, lecturer, department of life sciences BRACU for his invaluable insights, help, constructive criticism and support throughout the thesis process that immensely helped me broaden the spectrum of the work done on my thesis.
I thank Mr.Alamgir Hossain, computer lab in charge at MNS department for his constant support in regards of logistics and IT requirements.I thank all the staff of BRACU and my other well wishers for their help and support during my journey here as a student and beyond.

Rian Rafsan
Student ID: 21276004

Table of Contents

Chapter 1

Introduction

Elements with atomic numbers ranging from 57-71 form the lanthanide series in the periodic table. Also termed as rare earth metals (REMs), these elements usually occur within the earth's crust in the form of various primary and secondary minerals. Chemically inorganic phosphates, carbonates, fluorides or silicates, they develop via prolonged exposure to physical and chemical stresses over thousands and millions of years. Monazite, bastnaesite & xenotime are the usual ores used for lanthanide extraction.

Lanthanides are indeed of immense economic importance and a raw capital for many heavy industries. However, its economic and industrial importance often tends to overshadow the greater role it plays in the biological field-particularly within *in vivo* systems to be precise.Lanthanides were initially thought to be prone to biological inertia-typically rendering them virtually incapable of performing any biological function within the living system.

This grave misconception was rectified with the discovery of certain alcohol dehydrogenases (ADH) in proteobacterial species capable of utilizing the typical catalytic potential of lanthanides- the first of its kind to be discovered being *Methylacidiphilum fumariolicum* SolV (Pol,2014) which simultaneously showed capability and requirement of lanthanide dependent metabolism for faster & efficient breakdown of the metabolite PQQ in proteobacteria.

Dubbed as 'lanthanomes', these newly found novelties have been a subject of keen interest and meticulous research within the scientific community particularly in the last decade. These laborious efforts did not go futile and scientists soon found other proteobacterial species that bore further testimony of lanthanides' biological affinity. Furthermore, genes responsible for lanthanide dependent metabolism were soon identified- namely *xoxF*, *xoxG* & *xoxJ* respectively found to exist in an operonic manner.

While these events do prove that lanthanides are not ‘biologically dumb’ and are significant catalytic players of proteobacterial metabolism, the underlying cause of novelties among lanthanomes is yet to be properly understood.

A reason behind this is the high degree of heterogeneity among these novels who are typically ubiquitous by habitat and nutrition. As such, they don’t possess any sort of particular niche specificity- the sole features of their mutual homology being the proteobacterial ancestry and lanthanide dependency.

Secondly, while much effort has been given towards understanding lanthanomes, particularly in this decade, attempts were solely focused on an *in vitro* system with a pinch of phylogenetics added to the mix- resulting in a potential re-invention of the wheel with little to no new knowledge generated as a byproduct. While new lanthanomes were discovered in the process, these novel proteobacteria are still relatively a mystery to us.

Hence, in order to get a better understanding and subsequently reduce the existing knowledge gap regarding lanthanomes it’s essential to first and foremost reduce the degree of heterogeneity among these newly found novelties-particularly from a taxonomic perspective.

To address these issues, we designed our study around a rather non-conservative & unorthodox approach. Here, instead of looking for potential lanthanomes unlike in preceding studies, we rather conducted an *in silico* analysis of the bacterial pangenome on 31 non-redundant proteobacterial strains spanning 23 genera, all of which have been proven to be lanthanomes in an *in vitro* setup. A pangenome or supra genome refers to the entire set of genes present in the strains under a clade altogether. A pangenomic analysis further involves the classification of the genes identified into groups of core, soft-core, shell and cloud genes respectively based on the frequency of occurrence in the sample under consideration

Such a study can help us attain a better understanding of lanthanomes both from a genomic and taxonomic perspective alike. While we can get an idea of the genes crucial and unique to lanthanomes through this analysis, it can also essentially help us demystify these new found novels from a taxonomic perspective unraveling their salient features in the process.

Chapter 2
Literature Review

2.1: Lanthanides & the Advent of Lanthanomes:

Elements in the periodic table with atomic numbers ranging from 57-71 form the lanthanide series. Placed in group IIIB of the periodic table, these elements are also termed as ‘Rare Earth Metals’ due to their availability within the earth's crust in the form of minerals, who are quite naturally the ores used for economical extraction of these metals; monazite, pegmatite, bastnaesite and xenotime to name a few.

Despite the name, these metals are relatively abundant and are of massive importance from both industrial and economic perspectives alike. In fact, Cerium (58) is more abundant (68 parts per million) than copper (70 parts per million) and is the 25th most abundant element on the planet.

Lanthanides are relatively well known for their unusual versatility in heavy industries- for example, as a catalyst in the petroleum industry, as an X-ray intensifier, in the field of LASER dependent cinematography or even as a principal component in the development of hybrid alloys with hyper thermal-resistivity; ideally translating them as “Vitamins for the modern industry” (Balaram, 2019).

However, what truly sets these metals apart from other elements of the periodic table is their massive biological potential. In fact, the lanthanides were thought to be prone to extreme biological inertia until the recent decade; when lanthanide dependency was observed in various classes of proteobacteria for efficient metabolism, the first to be discovered being *Methylacidiphilum fumariolicum* SolV (Pol,2014).

Initially thought to be unique solely to methylotrophs, the hypothesis was proven wrong when lanthanide dependent alcohol dehydrogenation was observed in *Pseudomonas putida* KT2440 strain(Wehrmann,2017).

In 2018, a protein named lanmodulin was isolated and subsequently characterized from *Methylorubrum extorquens* AM1 (Cook, 2019) that is capable of selectively binding to lanthanides under *in vivo* conditions. (PDB ID: 6MI5)

2.2: Lanthanides-the biologically right choice:

The aforementioned events soon acted as a potential linchpin triggering massive interest within the global scientific community to study the biological potential of lanthanides ; subsequently the the word *lanthanome* was coined to signify any sort of biomolecule involved with the biochemistry of lanthanides.

As such, be it lanmodulin or the proteobacterial species mentioned earlier, all are inclusive of lanthanomes. In this study, however, our focus will be strictly be focused on bacterial lanthanomes, particularly from a pangenomic perspective. But before delving into lanthanomes, understanding the lanthanides is a prerequisite.

The newest entry to the bioorganic periodic table, lanthanides are f-block elements having the general electronic configuration of:

$$\mathbf{[Xe]4f^{1\text{-}14}5d^{0\text{-}1}6s^{2}}$$

Although, the biological potential of lanthanides did mesmerize the curious mind, rather than asking “Why choose lanthanides?” one should credibly ask “Why *not* should lanthanides be chosen?”- a question whose answer lies within its physical chemistry. (Migaszewski,2015,

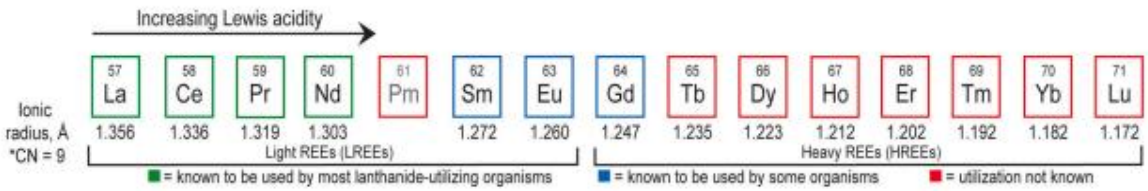

Figure 1: Periodicity of Lewis acidity & ionic radius within the lanthanide series.

The lanthanides, otherwise known as the rare-earths (REEs), are primarily categorized into 2 subgroups based on their individual atomic masses:

i)Light REEs (La-Eu)

ii)Heavy REEs (Gd-Lu)

Light REEs are relatively more abundant than their heavier counterparts; in particular, La-Nd has crustal abundance almost homologous to Cu & Zn. (Migaszewski 2015 , Cheisson 2019). Lanthanides are the biologically right choice due to their periodic properties of Lewis acidity & ionic radius respectively, both of which are periodic properties. In a period Lewis acidity increases when moves from left to right, subsequently decreasing its ionic radius as illustrated in Figure 1.

However, the lanthanides show an unusually higher degree of decrement in ionic radii than other periods and this unusuality decreases from Hafnium post lanthanide series. The unusual decrease being defined as **'Lanthanide contraction'**-a term coined by Victor Goldschmidt in his book entitled *Geochemische Verteilungsgesetze der Elemente* (Geochemical distribution laws of the elements).

Figure 2: Victor Goldschmidt, pioneer of lanthanide contraction.

Lanthanides are f-block elements; their contraction is simply a result of the poor shielding of nuclear charge(s) by 4f electrons leading to the 6s electrons being drawn towards the nucleus-significantly reducing the ionic radius in the process.

In single electron atoms, the average distance between electron & nucleus is determined by the subshell number decreasing with increasing nuclear charge-reducing the ionic radius in the process and aiding lanthanide contraction.

In the case of a multielectron atom, radial decrement comes with a nuclear charge increment partially offset by electrostatic repulsion among electrons. This is necessarily a shielding effect exerted by the inner electrons of the atom upon the electrons at higher energy levels decreasing in the order s>p>d>f.

But why make such a fuss about lanthanide contraction and why is it even important in lanthanide biochemistry? This rather novel property unique to lanthanides enables them with enhanced Lewis acidity and smaller than usual ionic radii rendering them biologically more

potent and efficient than previously discovered s, p or d block counterparts. In fact, despite the aqueous solubility of lanthanide ions being relatively low in near neutral pH (Firsching,1991), it's still significantly higher than Fe^{3+} which is well known for *in vivo* biocatalysis, particularly among bacteria, plants and fungal populations (Moeck,1998). Hence, lanthanides are essentially the 'biologically fit' candidates to be easily accessed *in vivo*.

Lanthanides, otherwise known as the 'inner transition elements' are unusually homologous to the d-block 'transition elements in terms of their physical and chemical properties alike. Elements of both categories tend to exhibit variable oxidation states & are capable of forming coloured and complex compounds by utilizing complex ligand molecules. Lanthanide contraction enhances the 'transition nature' of lanthanides by decreasing the ionic radii and increasing Lewis acidity of the lanthanides simultaneously, consequentially increasing their catalytic efficiency within the biological system in the process.

Lanthanides usually tend to exhibit the +III oxidation state with the sole exceptions of Ce and Eu with +IV & +II respectively. Furthermore, they have unusually high coordination numbers between 8 and 12. Which is the number of ions a central cation binds to form the complex. In most cases, the coordination number is typically 9. The high coordination number is also responsible for the selective behavior of lanthanomes towards lanthanides and as such the role of variability is also significant since a lower coordination number is preferred for LREEs while a higher number is preferred for the heavier ones.

All these features added together in lanthanides make them a highly efficient catalyst, particularly in the *in vivo* setup. Secondly, the relative abundance and ubiquity of the rare earths within the earth's crust also make them a highly economical option.

Thus, for the living system which is programmed with utmost subtlety and sophistication to run in the most probable efficient manner utilizing the least amount of resources required to avoid wastage, choosing lanthanides for biometabolism was not an event driven by chance but by need and itself being the best fit candidate to satisfy the need at that particular point of time.

2.3:Habit & Habitat Of Lanthanomes:

Contrary to most other members of the microbial community, lanthanomes still lack a complete taxonomic identity unique to these newfound novels. A reason behind this is the massive degree of heterogeneity amongst these newfound novels in respect of nutrition, habitat and genomic makeup alike.

From a taxonomic perspective, scientists report that lanthanomes tend to be prevalent amongst various classes of gram-negative proteobacteria; particularly in alpha, beta and gamma subclasses respectively(Daszczyńska, A 2022). Despite being highly prevalent within the *enterobacteriaceae* family, lanthanide dependent catalysis has also been reported in specimens belonging to families of *rhizobiaceae, pseudomonadaceae* & also *hyphomicrobiaceae* to name a few which are usually found in the soil. Methylotrophs are highly probable to be found in oceans and volcanic fissures whereas strains of lanthanomic *Pseudomonas* are typically free living. Thus both in respect of genomic diversity & habitat alike, these microorganisms are indeed highly heterogeneous. In terms of nutrition, they usually take 1 carbon, sometimes multi-carbon compounds as their typical food source.

The only points of intersection among these novel microbes are their ability to efficiently use lanthanides as catalysts for alcoholic dehydrogenation for concurrent efficient metabolism of PQQ, and their gram negative behavior. Still, the metabolic process is also different among different organisms, further enhancing their diversity in a multifarious manner.

2.4:Genes involved in lanthanide dependent metabolism in proteobacteria:

Lanthanides are typically responsible for the efficient catalysis of enzymatic dehydrogenation of alcohols in proteobacterial species, and quite a few genes are the key players cum regulators of this novel metabolic pathway.

Methanol dehydrogenase catalyzes the first step of methanol catabolism in methylotrophs and second step of methane conversion by methanotrophs respectively. Prior to the characterization of *xoxF* gene and the discovery of its lanthanide dependent catalytic activity, alcoholic dehydrogenation was presumed to be a process solely catalyzed by calcium using the *MxaFI* (otherwise called *Moxf*) gene. In both cases, however, PQQ acts as the catalytic center.

A salient feature of the genes involved in lanthanide dependent metabolism is that genes mostly tend to exist in an operonic manner. The first operon to be identified and found responsible for lanthanide dependent metabolism is the *xox* operon consisting of the following genes:

1) *xoxF*
2) *xoxG*
3) *xoxJ*

xoxF:

The first gene of the lanthanomic operon, *XOXF* is responsible for the encoding of methanol dehydrogenase. In the works of (Kawai, 2011(a),(b)) & (Nakagawa,2012), it was seminally proven that La^{+3} & Ce^{+3} are capable of inducing this gene in *Methylorubrum extorquens* AM1 and other methylotrophs. This subsequently leads to the characterization of the gene in *Methylacidiphilum fumariolicum* SolV (Pol,2014).

X-Ray Crystallographic comparison between active sites of both Calcium & lanthanide dependent MDHs showed a high degree of homology between the two proteins.

A point of distinction between the two is however, the presence of an extra aspartate residue in the lanthanide dependent moiety specifically at the 301^{st} position of the primary structure. Recent studies conducted on *xoxF1* isoform from *M extorquens* AM1 rendered the protein incapable of binding to lanthanide ions when the aforementioned aspartate residue.

Lanthanide dependent enzymes are one of those rare biomolecules that possess the liberty to be activated by multiple cations. However, despite the relative homologous chemistry among lanthanide cations, owing to lanthanide contraction attributes such as Lewis acidity, coordination number and respective ionic radii tend to differ and influence their enzymatic activity. Thus one of the major advantages of choosing lanthanides for efficient alcohol dehydrogenation besides being bioefficient is that there is a wider range of choices among

members of lanthanide series. Thus rendering Ln-dependent enzymes as one of the first multi-catalytic ones under the *in vivo* setup.

xoxG:

xoxG encodes a cytochrome-C that acts as the physiological electron acceptor for *XOXF* first reported to be discovered in *Methylomonas* sp. LW13(Chistoserdova, 2018). *XOXG* has also been characterized by *M extorquens* AM1 (Featherstone, 2019) and *M fumariolicum* (Versantvoort, 2019, Kalimuthu, 2019). In the case of *M extorquens* AM1, the protein of interest was characterized by crystallization and subsequent functionality was assessed using Nd, Ce & La *XOXF*s. It was observed that the metal in *XOXF* didn't affect the overall efficiency of *XOXF* but did impact the K_m of the expressed protein which is by definition, the substrate concentration necessary to reach half the maximum velocity of the enzyme (V_{max}).

xoxJ:

The third and last member of the *xox* operon that has a role to be played in the lanthanomic behavior of proteobacteria is the *xoxJ* gene that encodes a periplasmic protein associated with the **ATP Binding Cassette (ABC)** transport system. From a functional perspective, this periplasmic protein is presumed to act as a facilitator to enhance interaction between the MDH and cytochrome-C, meaning *xoxJ* is essentially the coordinator of the *xox* operon and a typical functional homologue of *MxaJ* gene of the calcium dependent MDH pathway. X-ray crystallography of *xoxJ* isolated from *M extorquens* AM1 revealed 2 globular domains flanking a putative ligand binding cavity larger than *mxaJ* & consists of a β-sheet missing several strands. Residues in the cavities are predominantly hydrophobic and bi-looped by the periphery. Hypothetically speaking, the hydrophobic cavity aids optimization of site specific binding for a large partially folded form of the *xoxF* placed under the same operon. This in particular is helpful for the development of *xoxF* apo-protein which is difficult to study in an activated form.

2.5:Other genes involved in lanthanide dependent metabolism;

As mentioned earlier, lanthanide dependent metabolism is independent of niche specificity, meaning that they are ubiquitous by nature and are not confined within a specific taxon or ecosystem.This heterogeneous behavior of lanthanomes has also been observed in their respective genomic make-up.

Before moving further, let us remind ourselves of the role of lanthanides. Lanthanides are essentially catalyzers of alcohol while the *xox* gene cluster is the first of its kind to exhibit lanthanide dependent catalysis in proteobacteria, it's not the only gene to exhibit lanthonomic behavior.

ExaF for example, is a gene that encodes a homodimer selective towards binding with La^{+3} over Ca^{+2} in its active site. While both *xox* and *ExaF* use PQQ as the common cofactor, *ExaF* is capable of oxidizing both ethanol and methanol, primarily ethanol, proving for the first time that lanthanides can catalyze reactions on multi-carbon substrates. V_{max} is nearly equal in both cases, however, K_m for ethanol in the case of *ExaF* is a bit lower than that of methanol. It also plays a secondary role in the oxidation of formaldehyde & acetaldehyde *in vitro*. *ExaF* was first isolated & characterized from *Methylorubrum extorquens* AM1 strain by observation of ADH activity in presence of LREEs despite the inactivation of every known ADH pathway including the *xox* operon (Vu, 2016). *ExaF* has been observed in strains of Beijerinckiaceae bacterial family members to exhibit lanthanide dependent catalysis. To be precise, the strains RHAL1, RHAL8 & RCH11 were reported positive for this particular gene as well as *xox* gene cluster by means of genome screening and PCR genotyping-with positive results being obtained for both the genes in both approaches mentioned here (Wegner, C.E. 2019).

The last gene to be characterized to have lanthanomic activity and the first such gene to be found outside the *methylobacteriaceae* family is the *PedH*. First reported to be observed in *Pseudomonas putida* KT2440 (Wegner, C E 2019), this gene encodes an ethanol dehydrogenase with a strong affinity towards Pm and Nd, although the *GOI* can bind up to Tb in the lanthanide series in an *in vitro* system. *In vivo*, La-Sm supports *PedH* activity if the calcium dependent counterpart *PedE* is deleted.

2.6:Lanthanide dependent metabolic pathways among Different Proteobacteria:

As mentioned earlier, lanthanide dependent catalysis is not an entity unique to a particular definite taxon; thus their heterogenous diversified behavior, although catalysis observed till now remains strictly confined within alcoholic dehydrogenation. This further implicates that lanthanide metals have the option to act on multiple gene products (proteins) responsible for multiple different pathways. This is further supportable by the poly-cystronic nature of prokaryotic genome. Thus lanthanomes of different taxa have lanthanides acting at different sites of the metabolic process. Here, we will take a look at the catalytic role of lanthanides in the metabolism of a few microbes belonging to different genera. This will further help us recognise the underlying genes in action and get a better understanding of the metabolic diversities in the process.

2.6.1: Lanthanide depend on metabolism in Methylotrophs:

Lanthanides have relatively poor water solubility and hence possess huge difficulties when it comes to their economical extraction from respective ores. As such, organisms that tend to utilize the catalytic efficiency of REEs should be able to secrete a lanthanophore, typically a protein for efficient intake of Ln species into the *in vivo* system. This selective and efficient uptake of the catalyst is aided by a TonB dependent system specific to lanthanides. Homologous systems have also been reported previously for copper and iron uptake (Daumann,2019). Studying *M extorquens* AM1, the *Lut* cluster was deemed responsible for encoding the TonB dependent system (Roszczenko-Jasińska et al., 2019) and is widely conserved among *Methylobacterium* genus members. *LutH* expresses the Ton-B dependent receptor, while *LutE* & *LutF* express ATPase & the membrane components of ABC transporters respectively. Mutants deficient of *LutH* failed to grow on methanol-Ln containing media, while *LutE* & *LutF* are inessential for lanthanomic activity in lanthanomes (Ochsner et al., 2019, Roszczenko-Jasińska et al., 2019). *xox*F is post translationally transferred from the cytosol and activated at the periplasm. *LutH* encodes the protein that transports lanthanides to the periplasm from the environment.

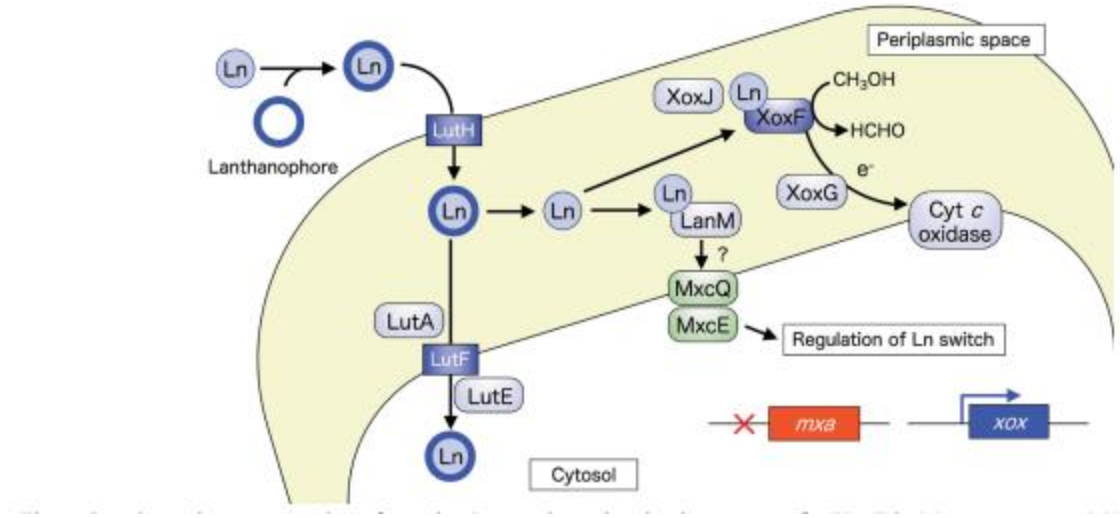

Figure 2: Lanthanophore bound to Ln is recognised by Ton-B receptor expressed by *LutH* gated with an N-terminal plug domain.Ln is released in periplasm activating *xoxF*.This in turn triggers methanol oxidation using *xoxG* and *xoxJ* genes of the operon who functionally encode the electron acceptor cytochrome-C and periplasmic binding protein for *xoxF* respectively.

2.6.2: Ln-dependent metabolism in rhizobia & *Bradyrhizobium*:

It was mentioned earlier that lanthanomes are typically ubiquitous by habitat and span multiple genera. Besides being abundantly found among members of *Methylobacteriaceae* family, they are also found in the plant symbiotic bacterial family of *Bradyrhizobiaceae* & *Rhizobiaceae* respectively.

There exist significant structural differences between the *xox* clusters of rhizobiaceae and *xox*1 cluster of methylobacteriaceae family members.

In rhizobia, the *xox* cluster is composed of 4 genes:

i) *xoxF*

ii) *xoxG*

iii) *fldA* expressing glutathione dependent formaldehyde dehydrogenase.

iv) *gfaA* expressing S-hydroxymethyl glutathione synthase.

In members of rhizobia, the complete methanol oxidation pathway is encoded by *xox*, despite formate dehydrogenase being located at a different genetic locus (Pastawan,2020). However, not all genes necessitated for methanol assimilation are present in rhizobia and methanol is likely to be a substep of many of the metabolic pathways yet to be discovered.

Besides methylotrophs, lanthanomes are hence also prevalent among plant symbionts. In the case of *Bradyrhizobium*, nodules are formed in the roots of leguminous plants where nitrogen is fixed to ammonia.*xoxF* is the dominantly expressing MDH in the case of *Bradyrhizobium*, with some exceptions (Pastawan,2020).

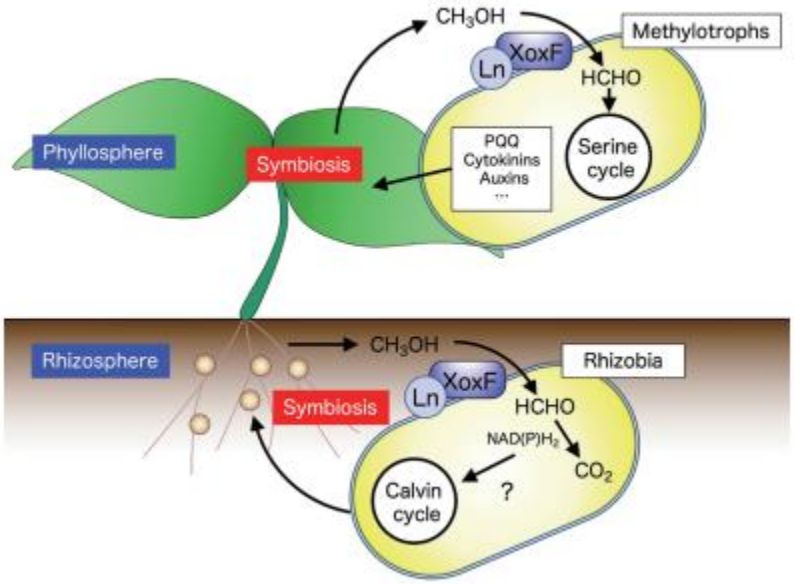

Figure 3: Plant symbiosis of Ln-dependent methylotrophs and rhizobia in phyllosphere & rhizosphere.

2.7: Regulation of Ln-dependent metabolism in proteobacteria:

Mechanisms underneath the efficient regulation of Ln-dependent catalysis in proteobacteria are still a subject of substantial diversity, although a unified theme amongst these multi-generic novels does exist.

Lanthanides are initially localized in the periplasm; signal is transduced to the cytosol for regulation via a bi-compartment system composed of a membrane-bound kinase sensor & a cytosolic response regulator.

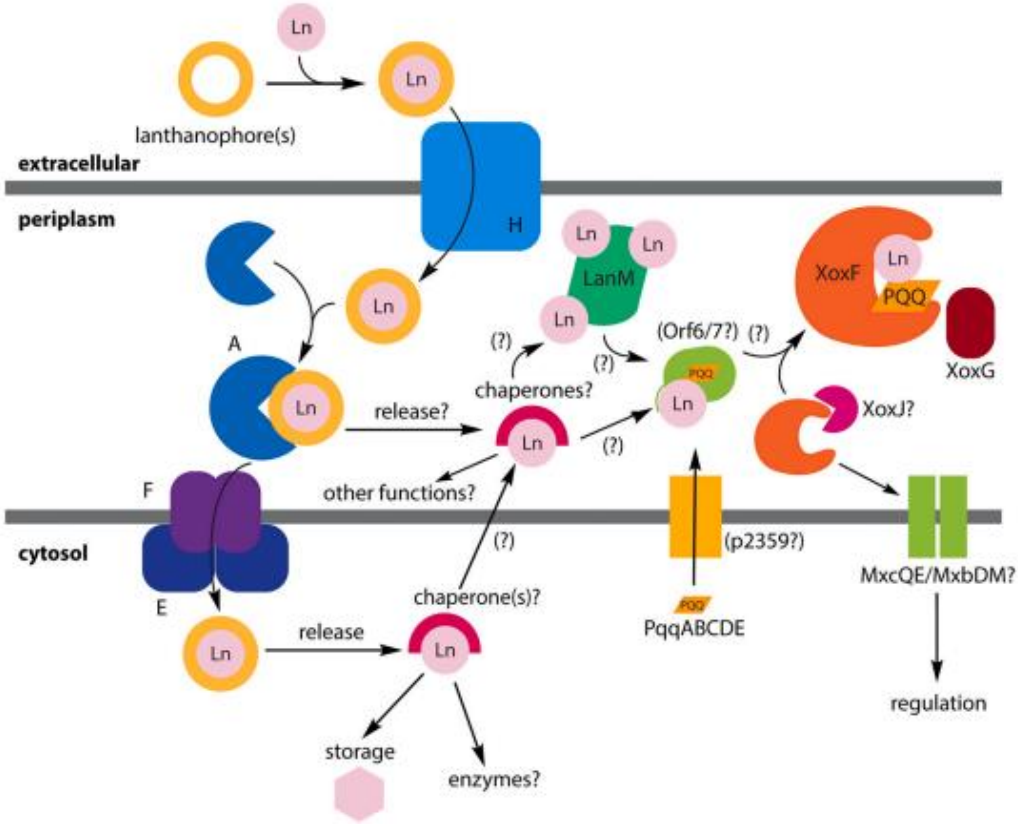

Figure 4: Model of lanthanide uptake, trafficking & subsequent utilization in *M extorquens*.

In organisms capable of utilizing both calcium and lanthanides for alcohol dehydrogenation, a 'Lanthanide switch' acts as the regulator. Generally, the presence of lanthanides *in vivo* induces the Ln-MDH transcription to repress the Ca-MDH. This regulation process varies among species. In the case of *M extorquens*, a two component system composed of MxbDM and MxcQE, along with response regulator MxaB, which is a coordinating protein. Another response regulator, namely MxbM is necessary for *xox1* in lanthanide's absence; *xoxF1* is also necessary for *MxaF1* expression & MxbDM system in addition to self repression. Hence, it's presumed that apo *xoxF* is in periplasmic Ln sensing utilizing MxcQ, MxbD or both; enzyme metalation to Ln(III) ion relieves *xox1* repression & subsequent *mxaFI* expression. In the case of *M buryatense*, MxbDM & MxcQE homologues are absent & sensor kinase MxaT unique to this species is present. For *P*

putida, *PedE* & *PedH* are the Ca-MDH & Ln-MDH encoding genes regulated by the PedR2/S2 system.

Chapter 3

Methods & Methodologies

The methodology applied in this work is a partial replication of the one used in (Gaba, S 2020) to study the pangenome and subsequently identify a superorder in halophiles for reconstruction of its ancestral state.

Correspondingly, in this study, we also conducted a pangenomic analysis of 31 non-redundant strains spanning 23 genera that were proven lanthanomic in an *in vitro* atmosphere. We conducted this analysis to get a better understanding of the genetic makeup of these new novel microbes. The step-wise protocol employed to satisfy this purpose is as follows:

3.1:Selection Of Lanthanomic Strains and Collection of Sequences to get the Required Dataset:

We began with the selection of strain-specific, gram negative & multigeneric proteobacteria that were previously proven to be capable of utilizing lanthanides under an *in vitro* setup .This particular behavior of utilizing lanthanides by biomolecules is lanthonomicity and strains exhibiting lanthonomicity are thus lanthanomic. Through meticulous and extensive literature review, we were able to select 31 non-redundant strains of lathanomic bacteria spanning 23 genera and the 3 classes of proteobacteria (now called *pseudomonadota*); namely alpha (α), beta (β) & gamma (γ) respectively.

Complete genomes of these 31 strains were collected; except for *Methylorubrum extorquens* AM1 which was collected from the NCBI genome database located at https://www.ncbi.nlm.nih.gov/genome, all other genomes were collected from the *Bacterial And Viral Bioinformatics Resource Center* database; https://www.bv-brc.org/ .All files were collected in DNA FASTA format.

3.2:Functional Annotation Of Collected Genomes:

The genomes collected earlier were subjected to the process of genome annotation; which refers to the process of identifying coding sequences present within a genome, followed by functional prediction of the protein(s) they express.

Genome annotation is dependent solely on the cellular structure of an organism; thus, the algorithm used for prokaryotic genome annotation is completely different from a eukaryotic one. The reason behind this is because the orf in a prokaryotic cell works in a polycistronic manner, unlike an eukaryotic cell that works in a monocistronic one.

Each genome in the sample was prokaryotic and was individually annotated using the *Rapid ProKaryotic Genome Annotation* software **(Prokka)** in short. Among the annotated files obtained as output for each organism, the GenBank format files (with .gbk file extension) were copied and pasted on a separate folder altogether, to conduct pan genomic analysis in the next step.

For annotation, DNA FASTA files were taken as input and were renamed such that no space was present in the name and the name did not start with any digit or specialized character. A typical command for annotating *Methylorubrum extorquens* AM1 in a machine with 7 cores will be as follows:

prokka —-cpus 7 —-kingdom Bacteria —-prefix Methylorubrum_extorquens_AM1 —-locustag Methylorubrum_Extorquens_AM1 Downloads/Methylorubrum_extorquens_AM1.fna

3.3: Clustering genes in Annotated Genome Sample Pre-Pan Genome Analysis:

Annotated genomes obtained earlier were next clustered in groups based on completeness and functional homology. Clustering in particular is crucial to the efficient classification of genes in the sampled organisms on the basis of frequency of prevalence.

The Get_Homologues software suite was the tool of choice for the identification of genetic clusters. Genetic clusters can be carried out in the Get_Homoogues software using one of the following 3 algorithms:

i)Bidirectional Best Hit (BDBH).
ii)Cluster Of Orthologous Sequences (COG).
iii)Orthologous Markov Clustering (OMCL).

Among these algorithms mentioned above, BDBH doesn't produce clusters against all BLAST driven approach. Rather, it chooses the smallest genome in the sample as a reference and keeps adding the rest of the genomes stepwise while storing the sequence clusters simultaneously margin the best hits in the process. While this approach is feasible for a core genome analysis, it's strictly non-applicable for a pangenomic study. The highly heterogeneous genomic structure and ubiquitous habitat of the lanthanomes add to the difficulty in this respect.

As such, we applied both OMCL & COG algorithms respectively to identify the gene and protein clusters present within the annotated genome sample. It should be mentioned prior that the primary focus of our work and subsequent results were developed around the OMCL. COG was later carried out to get a comparison with OMCL in terms of the number of gene and protein clusters obtained in each case.

Secondly, Prokka provides the annotated genome files in multiple file formats. Among them, the GenBank format files (with .gbk extensions) were used. This in particular is advantageous over DNA FASTA or protein FASTA files since it gives both gene clusters and protein clusters simultaneously at one go.

For the clustering steps, in both cases we had chosen a minimum coverage of 75% for pairwise alignment and the cut-off value was set 1e-10.

All annotated GenBank files were placed in a folder named 'annotated'; first COG and then OMCL algorithms were carried out by executing the following commands one after the completion of another.

```
get_homologues.pl -d annotated -G -t 0
```

For conducting OMCL, the following command was executed after completion of COG.

```
get_homologues.pl -d annotated/ -M -t 0 -c
```

It should be mentioned that the difference between OMCL and COG is that in the case of OMCL, the clustering is more uniform, and granularity is maintained due to control of the genetic inflation factor (F), which is the ratio between the median of the empirically observed distribution of test statistic to the expected median value. OMCL is thus more efficient, although computationally expensive.

In the case of COG, triangles of inter-genomic symmetric best hits are merged based on symmetry. In such cases, gene repetition among multiple clusters is a possibility.

Lastly, while conducting COG and OMCL, from a theoretical perspective, we should know the concepts of orthologues and im-paralogues.

Orthologous genes are genetic homologs within different organisms subject to genetic divergence due to speciation sharing a common genetic ancestry.

Paralogues are identical copies of the same gene present within an organism due to the process of genetic duplication.

3.4:Estimating the Lanthanomic Pangenome:

After the genome sample was subjected to the clustering algorithms, the pangenome matrix was formulated using the auxiliary perl script compare_clusters.pl accompanying the get_homologues software suite. The resulting outputs were respective text and phylip formatted files. The phylip format file was later used to get a graphical representation of the pan-genome tree using the **Integrated Tree Of Life (ITOL)** server located at https://itol.embl.de/upload.cgi .

The compare_clusters.pl script was also applicable to draw Venn diagrams of the gene clusters and protein clusters respectively as a comparison between COG & OMCL.

3.5:Plotting The Pan-genome Growth Curve & Core-Genome Decay Curve:

We further plotted a pan-genome growth curve and a core-genome decay curve for our genome sample. To plot these curves we used the plot_pancore_matrix.pl perl script after running OMCL and COG. The following commands were executed on terminal to plot the desired curves respectively.

```
plot_pancore_matrix.pl -i core_genome_algOMCL.tab -f core_both

plot_pancore_matrix.pl -i pan_genome_algOMCL.tab -f pan
```

The core-genome decay curve is the number of core-genes versus the genome curve, while the pangenome growth curve is the number of pan-genes versus genome growth curve. A salient feature of the core gene decay curve is that two different curves were plotted using two different curves fitting algorithms-Tittelin's and Willenbrock's respectively.

The pangenome growth curve was however plotted using Tittelin's exponential curve fitting equation.

3.6:Classification of Pangenomic Clusters on the Basis of Frequency Of Prevalence:

The pangenomic analysis conducted and in particular the clusters obtained as a byproduct of the OMCL algorithm are finally classified into 4 categories as follows depending on the frequency of prevalence within a sample under consideration.

Table 1: Gene Clusters present in a pangenome based on the frequency of prevalence.

Gene Category	**Frequency Of Prevalence within a sample**
Core Genes	Indispensable genetic element, present in 100% of the strains under consideration.
Soft-core Genes	Genes present in 95 % of the microbial sample under consideration, significantly essential to the organism.

Cloud Genes	Unique genes with highly strict strain specificity
Shell Genes	Remaining genes with moderate conservation rates, relatively dispensable

3.7: Searching Pangenomic Clusters For genes crucial to Lanthanomic behavior in proteobacterial sample:

The last step of our pangenomic study involved searching the genetic clusters obtained earlier from OMCL, for genes essential to lanthanide dependent metabolism in our proteobacterial sample.

Since genes responsible for lanthanide dependent metabolism were identified earlier *in vitro*, along with successful characterization of expressed protein being successfully done in most cases, searching the large & diverse lanthanomic gene pool using a pangenomic approach was relatively easier than expected.

In particular, we searched the gene pool for genes that are capable of utilizing the catalytic potential of lanthanides for the smooth functionality of their expressed proteins.

We did not confine ourselves there, we also tried to study these genes from an *in silico* framework- such as prediction of protein function by BLASTp. By searching for these genes within the available clusters we also tried to figure out their prevalence in the lanthanomic genomes and their potential genomic significance. Lastly, we fished for potential anomalies within the gene clusters and tried to explain them through critical scientific reasoning.

Chapter 4
Results & Discussion

4.1:Selection Of Strain Specific Multi-generic Proteobacteria on the basis of *in vitro* lanthanomic activity:

We started our study by selection of multi-generic, gram-negative proteobacterial strains proven to be lanthanomes under *in vitro* conditions. We chose these strains through meticulous literature reviews of works previously done on this topic, particularly from a genomic perspective (see references). These former studies had also deemed successful results concerning isolation & subsequent characterization of lanthanide dependent genes in most cases; a feat that will tend to prove advantageous to our work as we further delve into the results we achieved.

Our review process resulted in selection of 31 strain specific non-redundant lanthanomes spanning 23 specific genera altogether.

4.2:Analyzing the pangenome corresponding to our sample of Lanthanomic strains:

Strains presumed to be lanthanomes based on *in vitro* tests were analyzed through both COG and OMCL algorithms to identify orthologous gene and protein clusters for both algorithms. A pangenomic analysis typically produces the following outputs:

i) Gene clusters obtained by classification of gene pool on the basis of frequency of prevalence
ii) A core genes versus number of genomes plot
iii) A pan genes versus number of genomes plot.
Iv) A pan genome tree.

4.3:Distribution Of Gene Clusters Within the Pangenome based on Frequency Of Prevalence:

As mentioned earlier, a pangenomic analysis results in gene clusters searched in a sample on the basis of structural and functional homology, while subsequently clustered on the basis of

frequency of prevalence. The clusterwise gene distribution within our sample of lanthanomes is given in the pie chart below:

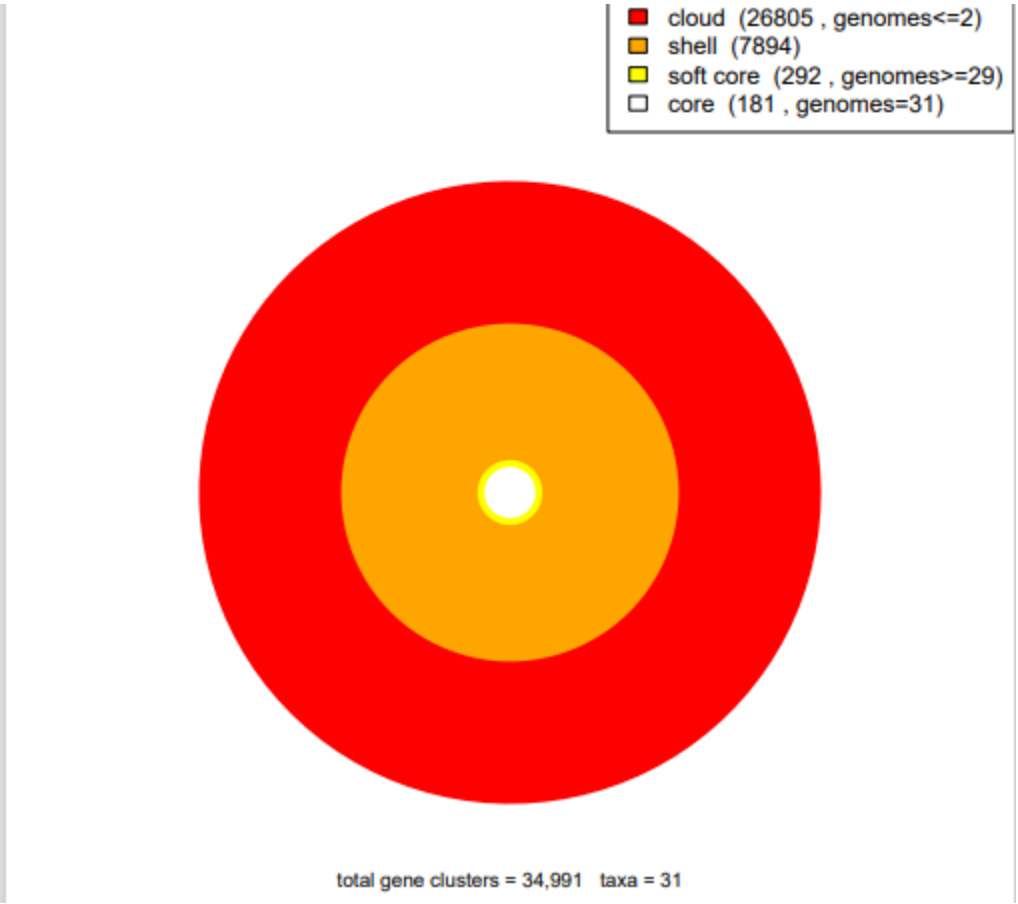

Figure 5: Pie chart depicting the distribution of gene clusters among sampled genomes.

As the pie-chart shows, the distribution of genes among sampled microbes is as follows:

Table 2: Number Of genes present in respective gene clusters with the corresponding number of strains in the sample.

Name Of Gene Cluster.	**Numbers Of genes**	**Number of Strains in sample with the genes.**
Core genes	181	31
Soft core genes	292	≥29
Shell genes	7894	Moderately conserved, relatively dispensable and variable by prevalence.
Cloud genes	26805	≤ 2
Total	34,991	

4.4:Selection Of Gene Targets to be studied *in silico* in microbial sample to verify Lanthanomic Activity:

Our next step towards reaching the solution space was the selection of suitable gene candidates to look for in the pangenomic pool who were previously reported to be lanthanomic under *in vitro* conditions. The following genes were chosen as credible gene targets to verify the lanthonomic activity of strains in our sample.

Table 3: List of genes screened for in this study with corresponding functions.

Gene Name	Function
xoxF	Encodes a Ln-dependent methanol dehydrogenase, part of *xox* operon.
xoxG	Encodes cytochrome-C that acts as electron acceptor for the *xox* gene cluster.
fldA	Encodes glutathione dependent formaldehyde dehydrogenase;part of *xox* gene cluster in lanthanomic nitrogen fixing symbionts.

gfaA	Encodes S-hydroxymethyl glutathione synthase, also part of *xox* cluster in nitrogen fixing lanthanomes.
ExaF	Encodes a homodimer selective towards binding with La^{+3} over Ca^{+2} and is capable of oxidizing both ethanol and methanol.
PedH	Encodes a Ln-dependent ethanol dehydrogenase first discovered in *Pseudomonas putida* KT2440.

4.5:Screening the pangenome for selected lantha-genes:

Pangenome obtained earlier as a byproduct of OMCL was then subjected to *in silico* screening for previously identified lanthanide dependent genes mentioned in the literature review.

Among the genes targeted to examine lanthanide dependency in our sample, only *fldA* and *gfaA* were recognised by get_homologues software using their respective trivial names.*xoxF* & *xoxG* on the other hand were recognised by the respective proteins they encode-namely methanol dehydrogenase and cytochrome-C to be precise.

xoxJ has not been completely characterized yet to identify the encoded protein. Hence, we did not choose the corresponding protein as a feasible gene target.

While *PedH* was identified as *qedA* by get_homologues suite and upon verification using BLASTp, it was confirmed that the gene was *PedH* and was encoding the protein lanthanide dependent ethanol dehydrogenase and hence is a typical lanthanomic gene. For example, we can show the BLASTp output for *Pseudomonas putida* GB1 *qedA* protein FASTA as follows:

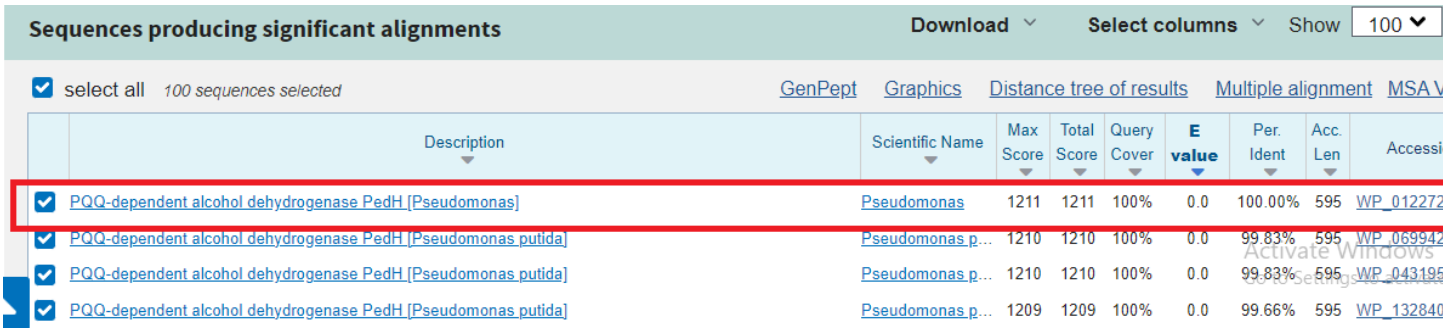

Figure 5:BLASTp result of *Pseudomonas putida* GB1 showing 100% homology for *PedH*.

ExaF on the other hand, was not recognisable by its trivial name and was erroneously identified as *ExaA*, a calcium dependent alcohol dehydrogenase but is actually a lanthanide dependent gene when checked in *Methylorubrum extorquens* AM1. Hence, to recognise *ExaF*, we took the corresponding protein sequence of *M extorquens* AM1 and carried out TBLASTn search for corresponding genomes matching them with our sample.

Results of our pangenome hunt for lanthanomic genes in our microbial sample are thus tabulated below as follows, along with references of proof of their *in vitro* lanthanide dependent activity :

Table 4: Tabulated result Of lanthanomic genes present in our sampled strains.

Name Of Strain	*xoxF*	*xoxG*	*PedH*	*ExaF*	*fldA*	*gfaA*	Reference of *in vitro* work where strain was mentioned lanthonomic
Advenella kashmirensis W13003		Positive				Positive	(Pyne,2018)

Beijerinckiaceae bacterium RH AL1	Positive(2 divergent gene copies)			Positive(Proven by TBLASTn)		Positive	(Wegner, 2019) PCR genotyping and genome screening.
Beijerinckiaceae bacterium RH AL8						Positive	(Wegner, 2019)
Beijerinckiaceae bacterium RH CH11						Positive	(Wegner, 2019)
Bradyrhizobium diazoefficiens USDA 110	Positive		Positive		Positive	Positive	(Pastawan, 2020)
Bradyrhizobium elkanii USDA76			Positive		Positive	Positive	(Pastawan, 2020)
Bradyrhizobium sp Ce3	Positive		Positive		Positive	Positive	(Pastawan, 2020)
Bradyrhizobium sp ORS278			Positive		Positive	Positive	(Pastawan, 2020)
Grimonti		Positive				Positive	(Huang et

a marina strain CECT 8713							al, 2019)
Methylacidiphilum fumarioli cum SolV							(Pol,2014)
Methylacidiphilum infernorum V4							(Huang et al, 2019)
Methylacidiphilum kamchatkense Kam1							(Huang et al, 2019)
Methylocapsa aurea strain KYG T		Positive					Positive for *xoxF* by genome screening in (Wegner, 2019).
Methylocella silvestris BL2	Positive (2 copies)	Positive					Positive for *xoxF* by genome screening in (Wegner, 2019).
Methylocella	Positive						Positive for *xoxF*

tundrae strain MTUNDRAET4 annotated genome							by genome screening in (Wegner, 2019).
Methylofe rula stellata AR4	Positive	Positive	Positive				Positive for *xoxF* by genome screening and PCR genoypin g in (Wegner, 2019).
Methylom icrobium buryatens e 5G							(Deng,20 18)
Methylom onas sp 11b	Positive						(Huang et al, 2019)
Methylom onas sp LW13							(Dauman n,2019)
Methylom onas sp MK1	Positive						(Huang et al, 2019)
Methylop ila M107	Positive						(Huang et al, 2019)
Methylor ubrum extorquen s AM1	Positive(2 non-divergent different			Positive (Proven by BLAST)			(Dauman n,2019)

	copies)						
Methylotenera mobilis 13	Positive(2 different copies,non-divergent)						(Daumann,2019)
Methyloversatilis sp FAM1			Positive(divergent in clusters 22952 & 108307)				(Huang et al, 2019)
Methyloversatilis universalis FAM5			Positive(divergent in clusters 22952 & 108307)				(Huang et al, 2019)
Methylovirgula ligni strain *BW863*	Positive (3 copies of the gene)						Positive for *xoxF* by genomic screening in Positive for *xoxF* by genome screening in (Wegner, 2019).
Pseudomonas			Positive(2 copies)				(Huang et al, 2019)

putida GB 1			(Detected by gene name & BLAST				
Pseudomonas putida KT2440			Positive(2 copies)				(Dauman n,2019), (Huang et al, 2019)
Pseudomonas putida SJTE 1			Positive(2 copies)				(Huang et al, 2019)
Sinorhizobium meliloti 5A14						Positive	(Huang, 2019)
Tistlia consotensis strain DSM 21585						Positive	(Huang et al, 2019)
Total	12	5	10	2	4	11	

4.6:Interpretation Of Results Obtained from Pan-genome search:

The pan-genome wide search was conducted in order to find lanthanomic gene candidates among our proteobacterial strain sample. Screening was conducted for specific gene targets. Genes found to possess at least one of the target genes were deemed lanthanomic in this study under *in silico* setup. Based on the criterion we thus fixed, 26 of the strains in our sample were proven to be capable of exhibiting lanthanide dependent catalysis through *in silico* pangenome analysis.

3 strains belonging to *Methylaciphilum* genus namely- *Methylacidiphilum fumariolicum* SolV, *Methylacidiphilum infernorum* V4 & *Methylacidiphilum kamchatkense* Kam1 were not found to possess any lanthanide dependent gene in their genome. Besides, *Methylomonas* sp LW13 & *Methylomicrobium buryatense* 5G did not show positive results for any of the markers we chose

to detect lanthanomicity in our strains. However, all the strains in our sample did exhibit *in vitro* lanthanide dependency as the references prove so.
Thus, mathematically speaking for a biased sample composed entirely of non-redundant strains proven to be lanthanomic under an *in vitro* setup, using an *in slico* pangenomic search driven approach we could prove that 83.8% of our samples were proven positive.

4.7: Possible Explanation Of Strains not being Identified & Other Findings:

5 of the strains in our sample failed to exhibit the presence of any of the gene candidates we had primarily chosen to confirm our sample's lanthonomicity in the pangenome analysis despite exhibiting lanthonomicity *in vitro*.

Why did such a problem occur? One possible answer is that our sample was highly heterogeneous owing to its multi-generic behavior and subsequently prone to genetic divergence. Hence, for the OMCL which clusters genes on the basis of functional homology and genetic distance between two neighborhood joining(NJ) genes, the distance was likely over the accepted threshold required for efficient execution of the algorithm, even though the algorithm is correct itself. Thus owing to high genetic divergence, the 5 strains may not have shown the desired genes in output although *in vitro* results deem them lanthanomic.

A second probable reason is that lanthanide dependence in proteobacterial metabolism itself shows variability among different taxa. For example, the genes *fldA* and *gfaA* are prevalent only to nitrogen fixing lanthanomes and not among members of *Methylobacteriaceae*. So the same gene, for example, *xox* works differently among different genera.

A third possibility is that not all of the genes under consideration are properly annotated within the database embedded in the get_homologues software in the first place. As a result, the genes might not have probably been found through our analysis.

The efficient execution of a program is always a trade off between time and accuracy. It is thus possible that running the analysis in better hardware with enhanced computational capabilities in terms of speed and accuracy may produce better results.

4.8:Other Novel Findings Of our Research:

Besides searching for novel lanthanomic genes in multigeneric proteobacteria, we also intended to study our sampled strain specimens from an *in silico* genetic perspective to get a better understanding of them. We did, in fact get to unearth some new findings regarding lanthanomes in this study as follows:

4.8.1:Where is the lanthanomic cluster?-Possible shell gene?:

First and foremost, the most common question while searching the pangenome for any gene is - *In what cluster is the Gene Of Interest (GOI) located based on frequency?*; such a question is usually asked to typically presume the significance of that particular gene in the genome of the corresponding organism.

In our case however, most of our genes searched for in the pangenome during research didn't belong to any cluster; with the sole exception of methanol dehydrogenase specific to *Beijerinckiaceae bacterium* RH AL1 being placed in strain-specific cloud gene cluster. While the gene is itself lanthanomic and upon BLASTp proved to be mutually isoforms of *xoxF*, we presume that this event typically proves the genetically divergent behavior of *xoxF*.

As mentioned earlier, one of our principal genetic identifiers for lanthanomic screening was the *xox* gene cluster. But, since it was a relatively new entry in the NCBI GenBank and was not completely annotated, along with the software itself containing a back-dated version of the database itself, searching for the corresponding protein expressed gave a better result that was further confirmed by running a BLASTp algorithm on the aa FASTA sequence.

While searching for the resultant methanol dehydrogenase did lead us to *xox* cluster- not all strains possessing *xox* cluster were initially identified. Upon critical scrutiny, it was observed that running the OMCL for pangenomic analysis on our sample had output the rare and unusual event of concurrent clustering.

The OMCL clustering algorithm clusters proteins into orthologous groups on the basis of functional homology and genetic distance. As mentioned earlier, alcoholic dehydrogenation in gram negative proteobacteria is not solely lanthanide dependent; a homologous metabolic pathway is also prevalent in these microbes which is calcium dependent, primarily regulated by the *moxF* cluster.

Upon examining the *moxF* cluster, we found proteins expressed by *xox* cluster, *PedH* gene as well as *ExaF* in the cluster file-all being lanthanomic and verified by BLASTp. Furthermore, *ExaA* and *ExaE* were also present in the cluster which are Ca-dependent nonetheless. One possible cause underlying this ambiguity is likely the multi-fate pathway of alcohol dehydrogenation dependent on the catalytic activity of different metals despite functional homology.

This brings us back to the actual question-where is the gene cluster? There is no specific gene cluster for lanthanide dependent alcohol dehydrogenation as per the obtained results-the operonic nature of *xox* may also have been responsible in this regard for algorithmic inefficiency.

However, if we apply the principle of exclusion-the lanthanide dependent alcohol dehydrogenases don't fall under either of the core or soft core genes since they are neither 95% nor 100% prevalent among the strains under consideration.

Had they been strain specific, they should have been placed under a cloud as in the case of *Beijerinckiaceae bacterium* RH AL1. However, since they don't belong to any of the three-the **shell cluster** is the likely option.

Owing to genetic divergence and relatively low conservation rates-strains capable of lanthanide dependent dehydrogenation also have the Ca-MDH pathway. Thus the lanthanide dependent genetic activity is not an indispensable metabolic event, but a better choice for efficient metabolism of PQQ as studies suggest.

Hence, considering all these facts the lanthanomic genes are likely members of the shell cluster, despite being subject to concurrent clustering due to algorithmic inefficiency.

4.8.2:*pqqB*-Core or soft-core?:

While lanthanide dependent alcohol dehydrogenation is simply a means to efficient metabolism of pyrroloquinoline quinone- *pqqB* being one of the essential genes responsible for its production, *pqqB* was observed both in the core and the soft-core clusters.

Upon meticulous scrutiny, it was observed that *pqqB*, like many other genes were prone to gene duplication when present in *Sinorhizobium meliloti* 5A14-a strain well known for gene duplication itself.

Thus since the gene was conserved yet duplicated in a strain within the sample, OMCL classified the gene as a member of both core and soft core clusters.

4.8.3: *PedH*-A lanthanide dependent ethanol dehydrogenase unique to *Pseudomonas*?:

While the *PedH* is a gene first characterized from *Pseudomonas*, pangenomic search and subsequent BLAST studies on functional homology studies proved the presence of this gene in strains of *Methyloversatilis* sp FAM1, *Methyloversatilis universalis* FAM5, *Bradyrhizobium diazoefficiens* USDA 110, *Bradyrhizobium elkanii* USDA76, *Bradyrhizobium* sp Ce3 & *Bradyrhizobium* sp ORS278 respectively. This gene, like *moxF* and *xoxF* is prone to genetic divergence despite being functionally conserved.

4.9:Venn Diagrams Depicting Number Of respective Gene Clusters & Protein Clusters Obtained through COG & OMCL:

We had run both the COG and OMCL algorithms over our annotated genome sets. We obtained gene clusters and subsequent protein clusters in both cases. Venn diagrams were constructed to get a comparative scenario of the respective gene and protein clusters for both algorithms used in our work.

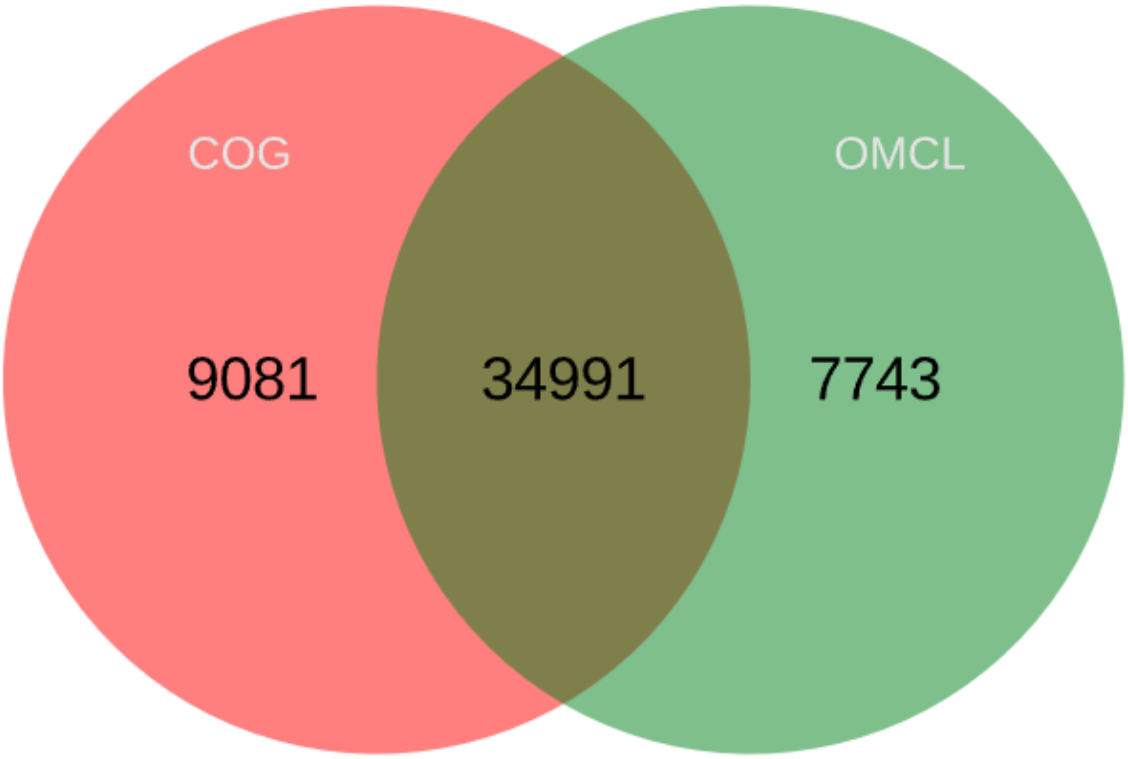

Figure 6: Venn diagram depicting the number of gene clusters using COG, OMCL & both algorithms.

Both COG & OMCL were successful in finding 34991 clusters mutually common, 9081 clusters were unique to COG and 7743 were unique to OMCL.

For translated protein clusters, the Venn diagram looked as follows:

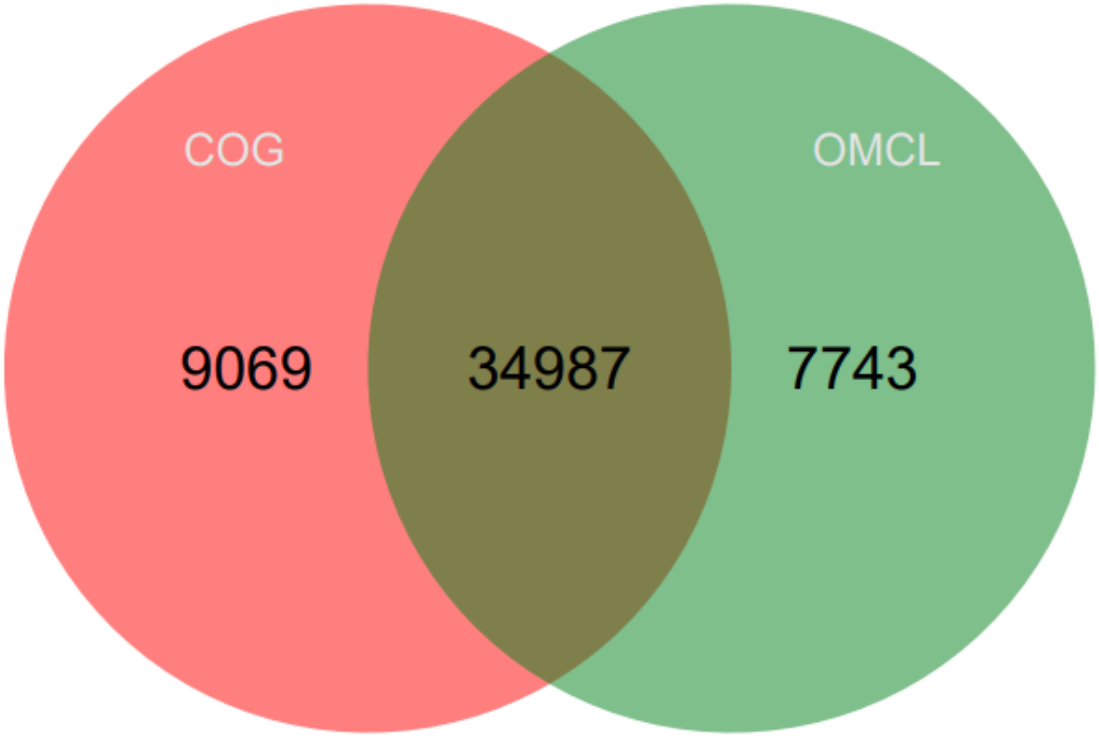

Figure 7: Clusters of expressed proteins depicted by Venn diagram for COG, OMCL & both algorithms.

COG resulted in 9069 unique clusters, OMCL resulted in 7743 unique protein clusters while 34987 clusters were identified by both the algorithms.

4.10: Core-genome Decay Curve:

A core-genome decay curve was plotted using both Tettelin curve fitting equation (in blue) and Willenbrock's curve fitting equation which were hard-coded within the software get_homologues. The curve thus obtained was plotted as follows:

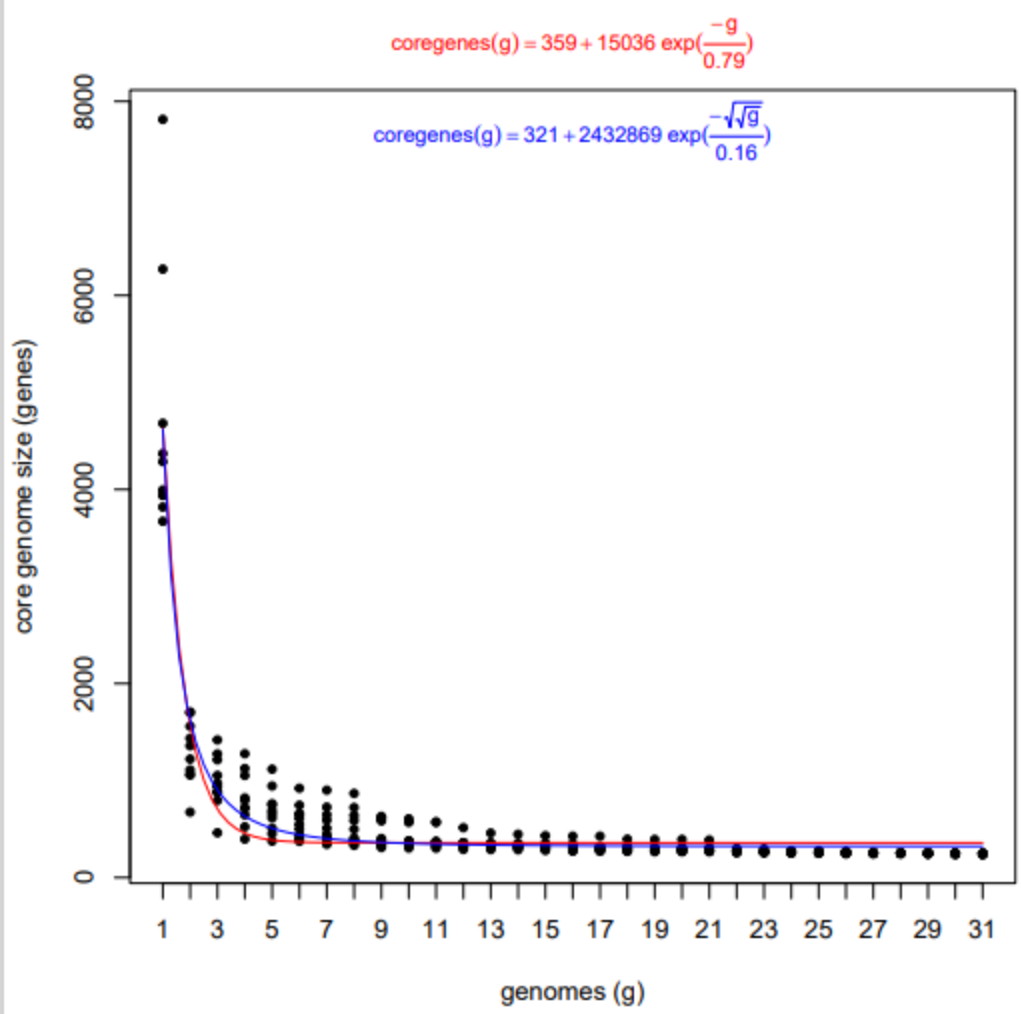

Figure 8: Core genome decay curve with both Willenbrock equation and Tettelin equation of curve fitting algorithm applied to our data set.

The strains in our sample are highly prone to genetic divergence with a low conservation rate showing fast decay. Hence our core genome curve is a closed one, meaning the addition of new genes with 100% conservation rate is highly unlikely.

4.11:Pan-genome growth curve:

A pangenome growth curve is plotted by plotting pan genes along the Y axis while the genomes are plotted over the X axis. The curve thus obtained following Willenbrock's parameters implicate that the lanthanomic pangenome is an open one and prone to horizontal gene transfer.

An open pangenome is one where the number of pan genes increases with the addition of new genomes to the pangenome set.

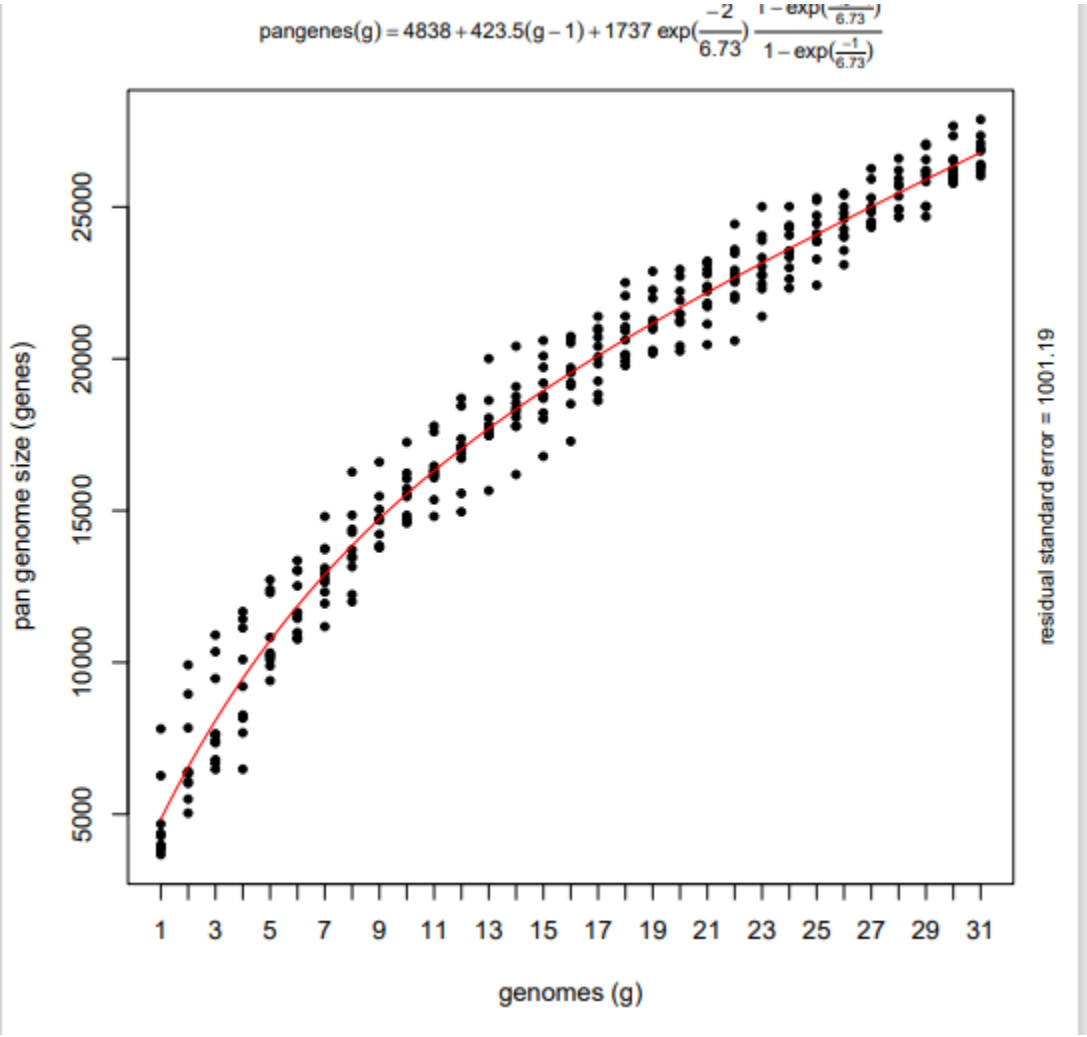

Figure 9: A pangenome growth curve showing the lanthanomes possessing an open pan genome.

4.12: Pangenome tree of multi-generic lanthanomes explained:

OMCL was also successful with respect to the successful construction of a pangenome tree. While trying to understand the phylogeny of a heterogeneous and diverse microbial group such as the lanthanomes, it's better to go for a multi-gene approach owing to its higher discriminatory power as opposed to a unigene approach such as the 16SrRNA metagenomic system.

A single gene approach would have been preferred if our concern was the study of housekeeping genes-where the core genes would have a crucial role to play. However, when the search is for

novelty within a highly divergent mixed microbial population, a pangenomic approach is the most accurate and efficient approach under optimum execution time to get the required output.

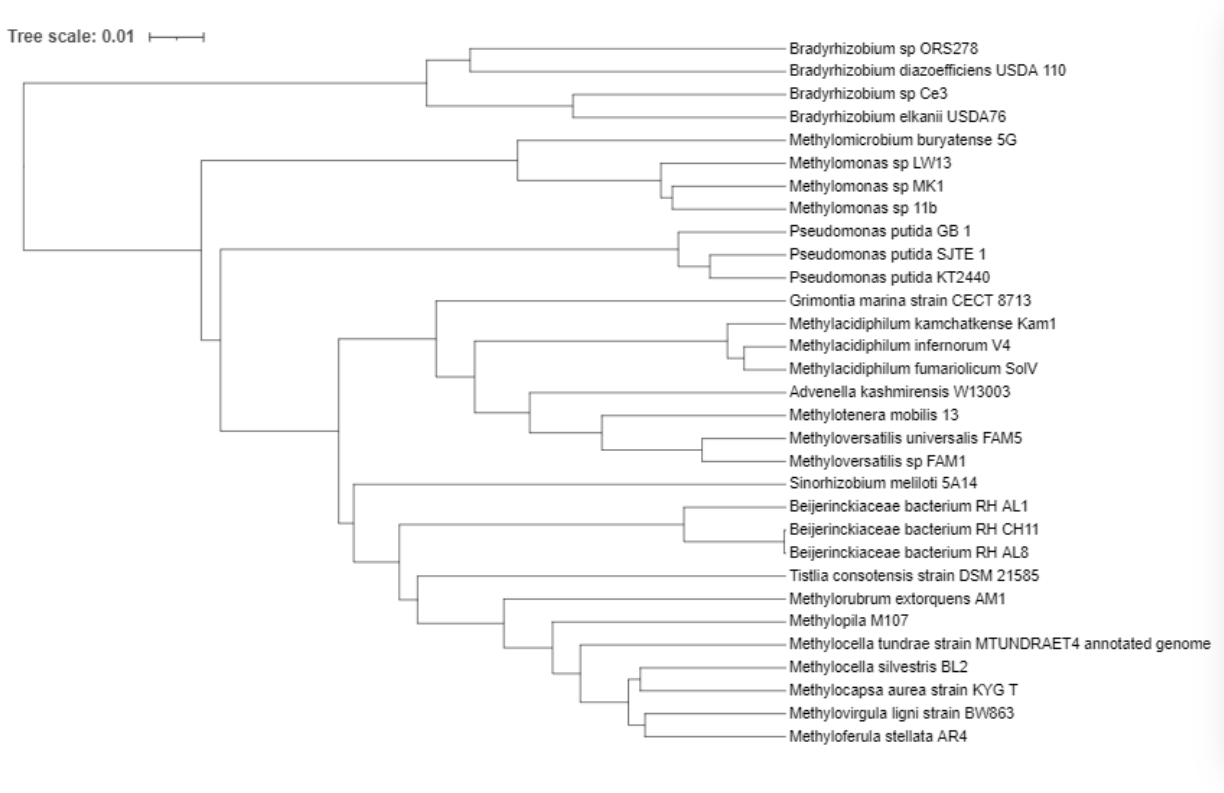

Figure 10: Pangenome tree with lanthanomic sample constructed with the ITOL server.

The pangenome tree exhibited a high prevalence of genetic divergence among lanthanomes. Nodes coming from the same root diverged along different directions as observed in the pangenomic tree.

For example, the strain of *Beijerinckia* bacterium RHAL1 diverged from its strains RHAL8 & RCH11 respectively and analysis proved it to possess two different functional isoforms of *xoxF* as mentioned earlier. *Methylocella silvestris* BL2 & *Methylocella tundra*e MTUNDRAET4 again although belonged to the same root diverged along different nodes over the course of time. These

are hence some instances of divergent behavior in these novels observed from a phylogenetic point of view.

Chapter 5
Future Prospects Of Research On Lanthanomes

5.1:What was it all about again?:

Here, in this research work we tried to conduct a pangenomic analysis on multi-generic strains on gram negative proteo bacterial species (now *pseudomonadota*) who exhibited lanthanide dependent catalytic activity if tested *in vitro*. Our approach towards this work was strictly restricted within the *in silico* paradigm.

We conducted such a work with the aim to attain a better understanding of these newly found novel bacteria that are diversified from multifarious dimensions-habit, habitat, nutrition, niche and metabolism simultaneously. This although may sound mesmerizing to the curious mind, but indeed proves to be a potential barrier to understanding these microbes-particularly from a taxonomic perspective.

We conducted this study as an attempt to technically reduce the potential void that exists within the knowledge space regarding lanthanomes. We believed that by applying a pangenomic approach on strains that were already proven to be lanthanomes, the resultant clusters that may be obtained through the analysis, will, in turn, help us understand the multigeneric nature of lanthanomes from a genomic and taxonomic viewpoint alike-which is still a challenge in this regard.

5.2:What results did our research produce again?:

Although our approach towards solving the existing problem in hand was an unorthodox one, it did give us results; results that nonetheless helped us get a better understanding of lanthanomes-as well as substantial numerical data to back our findings. We got to understand which genes are crucial to lanthanomes based on core gene cluster studies. We got to observe genetic divergence among these microbial elites, especially from a phylogenetic perspective,

Above all, we were able to identify potential genes that could be screened for in a microbial sample while searching for potential lanthanomes. Thus we believe, that our study on lanthanomes was quite successful-particularly if we consider the fact that ours was an *in silico*

approach to begin with-whereas most works done on these bacteria were limited to the *in vitro* set up.

5.3:Potential usage of lanthanomes-where to from here?:

As mentioned earlier, lanthanomes are a relatively new microbial group in the block. The fact that these microbes are highly efficient in utilizing lanthanides in their metabolic pathways added to the metal's economic value as an invaluable natural resource and versatile precursor to the heavy metal industry places lanthanomes in the central focus as the potential front-runners for the emerging biometallurgy industry in the near future. For example, harnessing the power of metabolic engineering and system biology, we can genetically modify lanthanomes to aid us in the cheap and economically feasible extraction of lanthanides from their respective ores (such as monazite, bastnaesite or xenotime to name a few) capitalizing on lanthanomes as potential bio-factories in this regard.

Figure 11: Pictures of monazite, bastnaesite and xenotime-ores of lanthanides used for extraction.

The industrial potential of lanthanomes is not limited to bio metallurgy, they are also usable through optimum genetic modifications as biological cleaners of lanthanide wastes in the environment. Lanthanides also form a huge bulk of medical waste worldwide -with approximately 20 tones of Gadolinium waste released as a byproduct of the MRI procedure in the US alone, which is a massive threat to the aquatic ecosystem already exposed to endangerment (Peplow, M. 2021).

Figure 12: Lanthanide wastes chunks waiting to be disposed off.

It doesn't end here; the discovery of lanmodulin (Cook,2019) marked a potential milestone in the development and discovery of lanthanomes. Until now, magnetosome was considered the ideal candidate for biomagnetism, which is being considered the future of biological hard-drives. But the discovery of lanmodulin and subsequently the credibility of lanthanides' potential in the development of ultra-hard magnets (Gould, C. A., 2022) does indeed show lanthanomes as a 'tiny giant' of the heavy metal industry in the very recent future.

5.4: How can our work help in terms of practical utilization of lanthanomes with enhanced efficiency?-What do we propose?:

Lanthanomes are highly heterogeneous in almost every respect; it's only their gram negative proteobacterial status and their ability to efficiently use the catalytic potential of lanthanide metals that connect them. So, how can we possibly utilize these novel microbes in such a way deemed beneficial for both the living world and the environment alike?

Before answering this question, we must remember that members of the lanthanide series and their derivatives possess massive industrial significance and typically contribute to modern heavy industries as 'essential vitamins'. But this also implies that they are equally responsible for contributing to the by-production of heavy industrial waste every year. Furthermore, these metals

are economically extracted from their ores- a process that is still quite costly in terms of the mass economics involved in the whole system.

Crude lanthanomes are only capable of utilizing lanthanides solely as catalysts. While the process in itself is amazing, it is naturally not possible to utilize it for an economic purpose- simply because the metal or metallic ore still remains unchanged *in situ*. Thus means should be adapted such that the lanthanomic potentials of these microorganisms are efficiently exploited to achieve our respective objectives- for example bio-metallurgy of lanthanides from ores and bioremediation of lanthanidous wastes from the environment.

How can these feats be possibly achieved? One possible way is through the means of metabolic engineering. Respective metabolic pathways that are lanthanide dependent have already been efficiently mapped and identified in lanthanomes, along with the genes involved and the proteins they encode. Furthermore, as our study has revealed, the principal metabolite of lanthanomes with respect to the lanthanomic pathway- PQQ-B, is a highly conserved and essentially a core protein to the lanthanome. Secondly, one of the major gene players in lanthanide dependent catalysis- *xox* is a gene cluster or operon bearing a tandemly repeated structure-making them easy to control and manipulate.

Thus, if we can metabolically engineer crude lanthanomes such that the metallic waste containing the lanthanides is consumed by the lanthanomes but not released unchanged as in a typical catalytic reaction, these microbes can potentially act as bioreactors for crude lanthanide extraction (when ore such as monazite is used as substrate), or carriers for metallic wastes which can later be disposed off through incineration.

This study can thus help identify suitable lanthanomic candidates who can be efficiently utilized for heavy bio-industrial purposes, through numeric and computational data obtained from an *in silico* pangenomic analysis.

References:

V. (2019). Rare earth elements: A review of applications, occurrence, exploration, analysis, recycling, and environmental impact. *Geoscience Frontiers*, *10*(4), 1285-1303.

Pol, A., Barends, T. R., Dietl, A., Khadem, A. F., Eygensteyn, J., Jetten, M. S., & Op den Camp, H. J. (2014). Rare earth metals are essential for methanotrophic life in volcanic mudpots. *Environmental microbiology*, *16*(1), 255-264.

Wehrmann, M., Billard, P., Martin-Meriadec, A., Zegeye, A., & Klebensberger, J. (2017). Functional role of lanthanides in enzymatic activity and transcriptional regulation of pyrroloquinoline quinone-dependent alcohol dehydrogenases in Pseudomonas putida KT2440. *MBio*, *8*(3), e00570-17.

Cook, E. C., Featherston, E. R., Showalter, S. A., & Cotruvo Jr, J. A. (2018). Structural basis for rare earth element recognition by Methylobacterium extorquens lanmodulin. *Biochemistry*, *58*(2), 120-125.

Migaszewski, Z. M., & Gałuszka, A. (2015). The characteristics, occurrence, and geochemical behavior of rare earth elements in the environment: a review. *Critical reviews in environmental science and technology*, *45*(5), 429-471.

Cheisson, T., & Schelter, E. J. (2019). Rare earth elements: Mendeleev's bane, modern marvels. *Science*, *363*(6426), 489-493.

Firsching, F. H., & Brune, S. N. (1991). Solubility products of the trivalent rare-earth phosphates. Journal of Chemical and Engineering Data, 36(1), 93-95.

Moeck, G. S., & Coulton, J. W. (1998). TonB-dependent iron acquisition: mechanisms of siderophore-mediated active transport. *Molecular microbiology*, *28*(4), 675-681.

Hibi, Y., Asai, K., Arafuka, H., Hamajima, M., Iwama, T., & Kawai, K. (2011). Molecular structure of La3+-induced methanol dehydrogenase-like protein in Methylobacterium radiotolerans. *Journal of bioscience and bioengineering*, *111*(5), 547-549.

Fitriyanto, N. A., Fushimi, M., Matsunaga, M., Pertiwiningrum, A., Iwama, T., & Kawai, K. (2011). Molecular structure and gene analysis of Ce3+-induced methanol dehydrogenase of Bradyrhizobium sp. MAFF211645. *Journal of bioscience and bioengineering*, *111*(6), 613-617.

Nakagawa, T., Mitsui, R., Tani, A., Sasa, K., Tashiro, S., Iwama, T., ... & Kawai, K. (2012). A catalytic role of XoxF1 as La3+-dependent methanol dehydrogenase in Methylobacterium extorquens strain AM1. *PloS one*, *7*(11), e50480.

Zheng, Y., Huang, J., Zhao, F., & Chistoserdova, L. (2018). Physiological effect of XoxG (4) on lanthanide-dependent methanotrophy. *MBio*, *9*(2), e02430-17.

Versantvoort, W., Pol, A., Daumann, L. J., Larrabee, J. A., Strayer, A. H., Jetten, M. S., ... & den Camp, H. J. O. (2019). Characterization of a novel cytochrome cGJ as the electron acceptor of XoxF-MDH in the thermoacidophilic methanotroph Methylacidiphilum fumariolicum SolV. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1867*(6), 595-603.

P. Kalimuthu, L.J. Daumann, A. Pol, H.J.M. Op den Camp, P.V. Bernhardt, Electrocatalysis of a europium-dependent bacterial methanol dehydrogenase with its physiological electron-acceptor cytochrome cGJ, Chem. Eur. J. 25 (2019) 8760–8768.

Vu, H. N., Subuyuj, G. A., Vijayakumar, S., Good, N. M., Martinez-Gomez, N. C., & Skovran, E. (2016). Lanthanide-dependent regulation of methanol oxidation systems in Methylobacterium extorquens AM1 and their contribution to methanol growth. *Journal of Bacteriology*, *198*(8), 1250-1259.

Wegner, C. E., Gorniak, L., Riedel, S., Westermann, M., & Küsel, K. (2019). Lanthanide-dependent methylotrophs of the family Beijerinckiaceae: physiological and genomic insights. *Applied and environmental microbiology*, *86*(1), e01830-19.

Daszczyńska, A., Krucoń, T., Stasiuk, R., Koblowska, M., & Matlakowska, R. (2022). Lanthanide-Dependent Methanol Metabolism of a Proteobacteria-Dominated Community in a Light Lanthanide-Rich Deep Environment. *International journal of molecular sciences*, *23*(7), 3947.

Daumann, L. J. (2019). Essential and ubiquitous: the emergence of lanthanide metallobiochemistry. *Angewandte Chemie International Edition*, *58*(37), 12795-12802.

Roszczenko-Jasińska, P., Vu, H. N., Subuyuj, G. A., Crisostomo, R. V., Cai, J., Lien, N. F., ... & Skovran, E. (2020). Gene products and processes contributing to lanthanide homeostasis and methanol metabolism in Methylorubrum extorquens AM1. *Scientific reports*, *10*(1), 1-15.

Ochsner, A. M., Hemmerle, L., Vonderach, T., Nüssli, R., Bortfeld-Miller, M., Hattendorf, B., & Vorholt, J. A. (2019). Use of rare-earth elements in the phyllosphere colonizer Methylobacterium extorquens PA1. *Molecular microbiology*, *111*(5), 1152-1166.

Pastawan, V., Suganuma, S., Mizuno, K., Wang, L., Tani, A., Mitsui, R., ... & Nakagawa, T. (2020). Regulation of lanthanide-dependent methanol oxidation pathway in the legume symbiotic nitrogen-fixing bacterium Bradyrhizobium sp. strain Ce-3. *Journal of bioscience and bioengineering*, *130*(6), 582-587.

Gaba, S., Kumari, A., Medema, M., & Kaushik, R. (2020). Pan-genome analysis and ancestral state reconstruction of class halobacteria: probability of a new super-order. *Scientific Reports*, *10*(1), 21205.

Tettelin, H., & Medini, D. (2020). The pangenome: Diversity, dynamics and evolution of genomes.

Peplow, M. (2021). Unlocking the lanthanome.

Gould, C. A., McClain, K. R., Reta, D., Kragskow, J. G., Marchiori, D. A., Lachman, E., ... & Long, J. R. (2022). Ultrahard magnetism from mixed-valence dilanthanide complexes with metal-metal bonding. *Science*, *375*(6577), 198-202.

Pyne, P., Alam, M., Rameez, M. J., Mandal, S., Sar, A., Mondal, N., ... & Ghosh, W. (2018). Homologs from sulfur oxidation (Sox) and methanol dehydrogenation (Xox) enzyme systems collaborate to give rise to a novel pathway of chemolithotrophic tetrathionate oxidation. *Molecular Microbiology*, *109*(2), 169-191.

Huang, J., Yu, Z., Groom, J., Cheng, J. F., Tarver, A., Yoshikuni, Y., & Chistoserdova, L. (2019). Rare earth element alcohol dehydrogenases widely occur among globally distributed, numerically abundant and environmentally important microbes. *The ISME Journal*, *13*(8), 2005-2017.

Wegner, C. E., Gorniak, L., Riedel, S., Westermann, M., & Küsel, K. (2019). Lanthanide-dependent methylotrophs of the family Beijerinckiaceae: physiological and genomic insights. *Applied and environmental microbiology*, *86*(1), e01830-19.

Deng, Y. W., Ro, S. Y., & Rosenzweig, A. C. (2018). Structure and function of the lanthanide-dependent methanol dehydrogenase XoxF from the methanotroph Methylomicrobium buryatense 5GB1C. *JBIC Journal of Biological Inorganic Chemistry*, *23*, 1037-1047.