

***IN SILICO* GENOME MINING FOR BIOSYNTHETIC GENE CLUSTERS FROM HALOPHILIC BACTERIA**

Submitted in partial fulfilment of the requirements for the award of Master of science in Bioinformatics and Data Science

By

NIVETHA R (41739014)



DEPARTMENT OF BIOINFORMATICS

SCHOOL OF BIO AND CHEMICAL ENGINEERING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

APRIL – 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC | 12B Status by UGC | Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Nivetha R (41739014)** who carried out the final year project entitled "***In Silico* Genome Mining for Biosynthetic Gene Clusters from Halophilic Bacteria**" under our supervision from January 2023 to May 2023.

Dr. JEMMY CHRISTY

Internal Guide

Dr. AMIT KUMAR

External Guide

Dr. JEMMY CHRISTY

Head of the Department

Submitted for Viva voce Examination held on

Internal Examiner

External Examiner

DECLARATION

I, **NIVETHA R (41739014)**, hereby declare that the Project Report entitled “***In Silico* Genome Mining for Biosynthetic Gene Clusters from Halophilic Bacteria**” done by me under the guidance of **Dr. AMIT KUMAR** and **Dr. JEMMY CHRISTY** and is submitted in partial fulfilment of the requirements for the award of Master of Science in Bioinformatics and Data Science.

DATE:

PLACE:

A handwritten signature in dark ink that reads "Nivetha" with a small flourish at the end.

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. JEMMY CHRISTY, Head of the Bioinformatics Department**, for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my External Project Guide **Dr. AMIT KUMAR** and Internal Project Guide **Dr. JEMMY CHRISTY** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Bioinformatics who were helpful in many ways for the completion of the project.

ABSTRACT

Halophilic bacteria have gained attention as a promising source of bioactive secondary metabolites with potential therapeutic and biotechnological applications. In this study, we conducted *in silico* genome mining to investigate potential biosynthetic gene clusters (BGCs) in 454 publicly available halophilic bacterial genomes. Our analysis revealed a diverse range of BGCs present across different clades and phyla, with some BGCs being more common than others, such as ectoine, terpene, betalactone, T3PKS, and RiPP-like BGCs. We also found that some species carried multiple copies of the same or different BGC classes, highlighting their potential for producing a variety of natural products. We also observed a moderate positive relationship between genome size and the number of BGCs present in halophilic bacterial genomes. Furthermore, our study highlights the potential of halophilic Actinobacteria as a rich source of BGCs and bioactive substances, particularly in the context of halophilic environments. The high diversity and abundance of BGCs in halophilic bacterial species suggest that these organisms could be a valuable resource for the discovery of new natural products with therapeutic and biotechnological applications.

Key words – Halophilic Bacteria, Natural products, Secondary metabolites, Genome Mining, Biosynthetic Gene Clusters, Therapeutic applications.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	i
	FIGURES	iv
	TABLES	v
	LIST OF ABBREVIATION	vi
1	INTRODUCTION	1
	1.1 Natural products	1
	1.2 Primary and Secondary Metabolites	3
	1.3 Microbial Natural Products	3
	1.4 Biosynthetic Gene Clusters	4
	1.5 Halophiles	5
	1.6 Halophilic Bacteria	6
2	REVIEW OF LITERATURE	8
	2.1 Salt - Loving Bacteria	8
	2.2 Bioprospecting	9
	2.3 Importance of Halophilic Bacteria	9
	2.4 Biosynthetic Gene Clusters and Classes	10
	2.5 Genomic Approaches	15
	2.6 Tools and Databases	17
	2.6.1 Description	18
3	AIM AND SCOPE	23
	3.1 Aim	23
	3.2 Scope	23
4	MATERIALS AND METHODS USED	24
	4.1 Data Collection using HaloDom	24
	4.1.1 Importing Data into Ms Excel	24
	4.1.1 Obtaining Genomic information	25
	4.2 Identification of BGCs	25
	4.2.1 antiSMASH	25
	4.2.2 antiSMASH Input Data	26
	4.2.3 Extra features of antiSMASH	26

	4.2.4 antiSMASH Output	28
	4.2.4.1 Manual Curation of Data into Excel	28
	4.3 Phylogenetic Tree Construction	29
	4.3.1 BioEdit	29
	4.3.2 MEGA	30
	4.4 Distribution of BGCs in Halophilic Bacteria	30
	4.4.1 iTOL	30
	4.4.2 Visualization and Annotation	31
5	RESULTS AND DISCUSSION	32
	5.1 Results	32
	5.1.1 Halophilic Bacterial Dataset Analysis	32
	5.1.2 Distribution of BGCs in Halophilic Bacteria	34
	5.1.3 Phylogenetic Diversity of BGCs in Halophilic Bacteria	38
	5.2 Discussion	41
	5.3 Future Scope	46
6	SUMMARY AND CONCLUSION	48
	REFERENCES	49

LIST OF FIGURES

FIGURE	DESCRIPTION	PAGE NO
5.1	Genomic availability of halophilic bacteria	32
5.2	Halotolerance classification of halophilic bacteria	33
5.3	Distribution of halotolerance and their genome completeness	34
5.4	Distribution of Biosynthetic Gene clusters (BGCs) in Halophilic bacteria	37
5.5	Correlation between genome size and BGC	38
5.6	Phylogenetic distribution of BGCs in Halophilic bacteria	39

LIST OF TABLES

TABLE	DESCRIPTION	PAGE NO
2.1	NRP and its bioactivity	12
2.2	Polyketide and its Bioactivity	13
2.3	RiPP and its bioactivity	14
2.4	Comparison of traditional natural product discovery and <i>in silico</i> genome mining of BGCs.	16
2.5	Tools and Databases	18
5.1	Top 20 Halophilic bacterial species contains more than 20 BGCs	35
5.2	Seven halophilic bacterial species did not contain BGCs	36
5.3	Genome Availability and Size of Halophilic Species Lacking Biosynthetic Gene Clusters (BGCs)	46

LIST OF ABBREVIATIONS

antiSMASH	Antibiotics and Secondary Metabolite Analysis SHell
BGC	Biosynthetic Gene Cluster
BiG-SCAPE	Biosynthetic Gene Similarity Clustering and Prospecting Engine
CDPS	tRNA-dependent cyclodipeptide synthases
CLUSEAN	CLuster SEquence Analyzer
DoBISCUIT	Database of BloSynthesis clusters CUrated and InTegrated
hglE-KS	Heterocyst glycolipid synthase-like PKS
IMG-ABC	Integrated Microbial Genomes Atlas of Biosynthetic gene Clusters
LAP	Linear azol(in)e-containing peptides
MIBiG	Minimum Information about a Biosynthetic Gene cluster
NAGGN	N-acetylglutaminyglutamine amide
NAPAA	Non-alpha poly-amino acids like e-Polylysine
NaPDoS	Natural Product Domain Seeker
NCBI	National Centre for Biotechnology Information
NRPS	Non-ribosomal peptide synthetase
PBDE	Polybrominated diphenyl ether
PKS	Polyketide Synthase
PUFA	Polyunsaturated fatty acid
RAST	Rapid Annotations using Subsystems Technology
RiPP	Ribosomally synthesized and post-translationally modified peptides
RRE-containing	RRE-element containing cluster
SBSPKS	Self-Organizing Map-Based Polyketide Synthase Prediction System
T1PKS	Type I PKS
T2PKS	Type II PKS
T3PKS	Type III PKS

CHAPTER 1

INTRODUCTION

1.1 Natural Products

Natural products are small chemical compounds that are produced by living organisms, such as plants, animals, and microorganisms, and have been used for centuries in traditional medicine. These compounds are usually synthesized by the organism for its own purposes, such as defense against predators, communication or to attract pollinators. They can have a unique and structurally diverse chemical structures and biological activities, and they have been used in medicine, agriculture and other industries for centuries. Essential oils, alkaloids, flavonoids, terpenes, and phenolic compounds are some examples of natural products (Atanasov *et al.*, 2021).

Natural products have historically been vital to the discovery of new therapeutic candidates and continue to be so today. Natural products are extremely complex and challenging to synthesize, which can be both a disadvantage and a benefit. Natural products' complexity can make them challenging to research and modify, but it can also imply that they may have various mechanisms of action, which can be useful for treating severe disorders (Dias *et al.*, 2012).

These compounds have acquired a great deal of attention in recent years due to their potential to be a rich source of bioactive chemicals. Bioactive molecules are substances that have the potential to be utilized as drugs or therapeutic agents because they have biological activity (Atanasov *et al.*, 2021).

The conventional way of isolating bioactive compounds from natural products involves biological screening of crude extracts to identify a bioactive hit extract, which is then further fractionated to isolate the active natural products. It involves variety of techniques such as solvent extraction, chromatography and crystallization. This process is known as bioactivity-guided isolation (Sasidharan *et al.*, 2010).

Bioactivity-guided isolation has several limitations

1. It is a laborious process that requires several chromatographic techniques and various other purification methods to isolate bioactive compounds.
2. It is a time-consuming process that can take several months to identify a single bioactive compound.

On the other hand, genome-based approaches have been developed in recent years to predict bioactive compounds. These approaches rely on the identification of biosynthetic gene clusters (BGCs) in genomes, which encode the enzymes and regulatory elements necessary for the production of natural products. These BGCs can then be expressed and characterized using synthetic biology techniques, allowing researchers to identify and study the bioactive molecules they produce. These approaches have the advantage of being faster and more efficient approach than traditional methods of isolating bioactive compounds (Bauman *et al.*, 2021).

One advantage of genome-based approach is that they allow researchers to analyze the biosynthesis of bioactive compounds at the genetic level. This enables a more detailed knowledge of the biochemical processes involved in the synthesis of these compounds. By studying the regulation of these pathways, researchers can identify ways to optimize the production of bioactive molecules or to produce novel analogs with improved properties. It is a promising method for discovering new bioactive compounds (Atanasov *et al.*, 2021).

Sequencing technology has become cheaper and faster over the years due to advances in technology and improvements in sequencing methods. This made it more accessible and affordable for researchers to sequence genomes, transcriptomes, and metagenomes and study the genetic basis of various biological processes. The number of sequences available in public databases has increased significantly in recent years. The availability of many genome sequences in public databases can be extremely helpful for genome mining. Genome mining involves searching for genes and gene clusters that are involved in the biosynthesis of natural products, such as antibiotics and other bioactive compounds (Kautsar *et al.*, 2020).

However, the sheer volume of sequence data can be overwhelming, and it has also created new challenges in terms of data analysis and interpretation. To address this

challenge, various tools and databases have been developed to help researchers navigate and analyze sequence data.

For example, NCBI GenBank database contains over 200 million sequences, and provides tools for searching, downloading, and analyzing sequence data. Other databases, such as antiSMASH-DB and Atlas of Biosynthetic Gene Clusters, are specifically designed for genome mining and provide information on biosynthetic gene clusters and natural product biosynthesis pathways.

1.2 Primary and Secondary metabolites

Primary metabolites are compounds that directly contribute to an organism's metabolic activities and are required for appropriate growth, development, and reproduction. While secondary metabolites are not necessary for the growth of organisms, yet they are essential for their survival and environmental adaptability, for example by acting as defensive compounds or signaling molecules in ecological processes communication. Examples of primary metabolites include nucleotides, amino acids, and carbohydrates; these compounds play a significant role in primary biological functions such as DNA replication, protein synthesis, and energy production. Due to their wide structural makeup and biological activity, a number of secondary metabolites have been used in agriculture, medicine, and other fields. They are present in all living things, including bacteria, plants, and mammals. The study of secondary metabolites contributed to the development of drugs to treat of human infectious diseases. Herbal plants, invertebrate animals, and microorganisms like fungi, algae, bacteria, actinobacteria, and cyanobacteria attracted more interest in research. The enormous diversity of secondary metabolites' structure, function, and biosynthesis can be used to classify them. There are five primary categories of secondary metabolites: terpenoids and steroids, fatty acid-derived compounds and polyketides, alkaloids, non-ribosomal polyketides, and enzyme cofactors (Vijayakumar & S. Raja, 2018).

1.3 Microbial Natural Products

The leading producers of natural products are microorganisms. Microorganisms, which include bacteria, fungus, and protists, are immensely diverse. Because of their unusual structures and functions, microbial natural products are of great importance. These substances (secondary metabolites) are widely produced

through specialized metabolic processes that are different from primary metabolic processes engaged in fundamental cellular functions. Secondary metabolites are organic compounds of low molecular mass. They serve various survival functions for the organisms producing them. These metabolites have a wide range of metabolic origins and unique chemical structures. And they act as the basis for various important industrial, agricultural, and medical products. As a natural source for pharmaceuticals, pesticides, and other chemicals, they are of great significance on human life. The activities and functions of secondary metabolites are not essential for a microbe's survival, but they may give the organism an advantage in the face of environmental pressures. The ability to produce secondary metabolites are encoded by groups of genes called biosynthetic gene clusters (BGCs) (Demain, 2014).

1.4 Biosynthetic Gene Clusters

Biosynthetic gene clusters are clusters of non-homologous genes that are physically close to one another on the genome, are tightly linked, and participate in the common metabolic pathway. Secondary metabolites are a class of chemical molecules with potential medicinal uses that are produced by biosynthetic gene clusters. Computational analysis of biosynthetic gene clusters has enhanced the natural product discovery by enabling the rapid investigation of secondary metabolic potential within microbial genome sequences (Kautsar *et al.*, 2020).

One of the most commonly used tools for genome mining of BGCs is antiSMASH (Antibiotics and Secondary Metabolite Analysis SHell). antiSMASH is a web-based tool that can identify and annotate BGCs in bacterial genomes. It uses a combination of algorithms to predict the structure and function of the biosynthetic gene clusters, as well as the potential bioactive molecules that they produce.

In addition to antiSMASH, there are several other bioinformatics tools that can be used for predicting biosynthetic gene clusters (BGCs), including MIBiG (Minimum Information about a Biosynthetic Gene cluster) and PRISM (Prediction Informatics for Secondary Metabolomes). These resources are crucial for the identification and evaluation of secondary metabolite biosynthetic gene clusters, which are in

responsible of producing a range of bioactive substances with potential medicinal uses (Medema *et al.*, 2015).

Natural product discovery and drug development program now immensely rely on genome mining for biosynthetic gene clusters in microbial genomes.

1.5 Halophiles

Halophilic organisms are extremophiles that thrive in environments with high salt concentrations. These organisms can be found in hypersaline environments such as salt lakes, salt pans, and salted food products. Halophilic organisms have unique metabolic pathways that have potential applications in agriculture, industry, and medicine. They are recognized producers of carotenoid pigments, retinal proteins, hydrolytic enzymes, and compatible solutes. They are found in all three domains of life: Archaea, Bacteria, and Eukarya. They are classified into three groups based on their degree of salt requirements: slight halophiles (0.34-0.85 M salt), moderate halophiles (0.85-3.4 M salt), and extreme halophiles (3.4-5.1 M salt) (Corral *et al.*, 2019).

Most halophilic organisms exclude salt from their cytoplasm by expending energy to avoid protein aggregation or “salting out”. This is known as high-salt-in-strategy. Halophilic organisms have developed various strategies to cope with the high concentrations of salt in their environment. One method is to accumulate compatible solutes, which are small organic molecules that help to maintain the osmotic balance between the cell and the environment. Ectoine, glycine betaine, trehalose are some common compatible solutes. Another method is to modify the cell membrane to make it more resistant to salt. These adaptations allow halophilic organisms to thrive in environments with salt concentrations that are inhospitable to most other organisms. Halophilic organisms have attracted attention due to their ability to grow under extreme conditions make them a valuable resource for biotechnological applications and the discovery of novel bioactive compounds with potential therapeutic applications (Dutta & Bandopadhyay, 2022).

There are several halophile databases available online, including HaloWeb, HaloBase, HProtDB, and HaloDom. Overall, HaloDom is a comprehensive database that covers halophiles across all life domains and provides valuable information for research in various fields (Loukas *et al.*, 2018).

Researchers are studying the biosynthetic gene clusters of halophilic organisms to identify and characterize the genes responsible for producing these specialized metabolites. The study of halophilic organisms and their BGCs has led to the discovery of new, potentially valuable secondary metabolites. These metabolites have a wide range of applications, including the development of new antibiotics, antifungals, and anticancer drugs (Corral *et al.*, 2019).

1.6 Halophilic Bacteria

In hypersaline habitats, halophilic bacteria are the most prevalent microbes, and as a result, they play an important role in maintaining the ecological balance of these environments (Edbeib *et al.*, 2016).

Halophilic bacteria in particular are known to produce bioactive substances, enzymes, and secondary metabolites with pharmaceutical applications among the various halophiles. Researchers are interested in these bacteria because they have particular adaptations to thrive in hypersaline environments.

Unlike most other microorganisms, halophilic bacteria have special biochemical and physiological adaptations that enable them to endure at high salt concentrations. As a result, they are capable of producing a variety of bioactive substances and have a wide range of potential biotechnological uses, including bioremediation, medicines, and food production.

Halophilic bacteria are an attractive target for drug discovery due to their unique capacities and tolerances. Recent studies have shown that metabolites from halophilic bacteria have potential applications in medicine, especially in cancer treatment (Mohammadipanah *et al.*, 2015).

Therefore, this study focusses on *in silico* genome mining of biosynthetic gene clusters in halophilic bacteria. Halophilic bacteria are specific microorganisms that thrive in high salt environments, and they have developed specific mechanisms to adapt to these extreme conditions. These mechanisms include the synthesis of specialized metabolites, many of which are bioactive and have potential application in biotechnology, agriculture, and pharmaceuticals (Dutta & Bandopadhyay, 2022).

Genome-based approaches, like *In silico* genome mining, have revolutionized the search for new bioactive compounds and improved the study of natural products (Atanasov *et al.*, 2021).

In this project, we will explore the diversity of biosynthetic gene clusters in halophilic bacteria, and our project aims to uncover novel and exciting bioactive compounds that have the potential to revolutionize the field of drug discovery and biotechnology, paving the way for new and innovative solutions to pressing global challenges.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Salt - Loving Bacteria

Halophiles are organisms that thrive in high salt concentrations and are found in areas where the concentration of salt is five times greater than salt concentration of the ocean, such as the Great Salt Lake in Utah, Owens Lake in California, the Dead Sea, and in evaporation ponds. Halophiles can be classified into the domains Archaea, Bacteria, and Eukarya, and they may be described as slightly, moderately, or extremely halophilic based on the extent of their tolerance to salinity. Slight halophiles prefer (1-3% NaCl) moderate halophiles (3-15% NaCl) and extreme halophiles (15-30% NaCl). Halophiles have unique biomolecules that have potential applications in biomedicine (Azhar *et al.*, 2014).

The general features of halophilic microorganisms are the low nutritional requirements and resistance to high concentrations of salt with the capacity to balance the osmotic pressure of the environment. They are physiologically diverse, and they can be aerobic or anaerobic, heterotrophic, phototrophic, and chemoautotrophic. Ecologically, the halophilic microorganisms inhabit different ecosystems characterized by a salinity higher than seawater, such as hypersaline soils, springs, salt lakes, sabkhas, and other naturally-occurring coastal saline habitats, marshes, marine abyssal sediments, and endophytes. Other known habitats are the result of human intervention like salted foods, brines, oil fields, saltern ponds and tanneries. These habitats are extreme and possess limited microbial diversity because of the combined effects of high salinity, high temperature, and high radiation (Corral *et al.*, 2019).

Halophiles have adapted to these hypersaline environments by adopting various strategies to compensate for their saline surroundings. Most halophilic organisms cope with the high concentrations of salt by expending energy to exclude salt from their cytoplasm. They prevent the loss of water by synthesizing or accumulating compatible solutes, which are small organic molecules that do not interfere with cellular functions and help maintain the osmotic balance. Some halophiles have also evolved specialized enzymes and transporters that function optimally in high salt

concentrations, and they have developed mechanisms to protect their DNA from damage caused by high radiation levels. These adaptations make halophiles a valuable resource for biotechnology and bioremediation, as they can be used to produce enzymes and biomolecules that are stable in high salt concentrations (Edbeib *et al.*, 2016).

2.2 Bioprospecting

The exploitation of extremophiles is having special importance in the development of new molecules with potential applications in biomedicine. Extremophiles, including halophiles have evolved to survive in extreme environments, and as a result, they have developed unique adaptations and produce bioactive compounds that may have therapeutic applications. Halophiles produce bioactive compounds such as antibiotics, antivirals, and anticancer agents, that have potential therapeutic applications. Halocins, which are natural antimicrobial peptides produced by halophilic archaea and bacteria, have potential applications in biotechnology and medicine as natural antimicrobial agents. The exploitation of halophiles is an area of active research, and scientists are continually exploring new ways to harness the unique properties of these organisms for various applications. The production of bioactive compounds by halophiles is a promising area of research, and ongoing studies are exploring the biosynthesis and potential applications of these compounds (Corral *et al.*, 2019).

2.3 Importance of Halophilic Bacteria

Actinobacteria, particularly the genera *Streptomyces* and *Nocardiosis*, are responsible for inhibiting clinical pathogens. These bacteria are frequently isolated from solar salterns, sea floor sediments, and mangroves (Xu *et al.*, 2018). Actinobacteria are known for producing a variety of natural products, including enzymes, enzyme inhibitors, signaling molecules, and immunomodulators, which have potential applications in medicine and biotechnology (Barka *et al.*, 2016).

Proteobacteria is the most frequently isolated group from marine environments when using plating techniques, and the genera *Virgibacillus* and *Vibrio* are largely recognized as halophilic. Some halophilic bacterial isolates of the genera *Halobacillus*, *Sanilivibrio*, *Oceanobacillus*, *Salinicoccus* have shown potential

antimicrobial activity against *E. coli*, *S. aureus*, *P. aeruginosa*, and *Salmonella typhi* (Conde-Martínez *et al.*, 2017).

Halophilic cyanobacteria produce large amounts of compatible solutes, which are organic compounds that help them to survive in high salt concentrations. Some studies have shown that halophilic cyanobacteria can produce bioactive metabolites with diverse biological activities, such as antiviral, antibacterial, antifungal and anticancer (Abed *et al.*, 2009).

2.4 Biosynthetic Gene Clusters and Classes

Biosynthetic gene clusters (BGCs) are groups of genes that are responsible for producing specialized metabolites, such as antibiotics, anti-cancer agents, and other bioactive compounds. These clusters can be found in a variety of organisms, including bacteria, fungi, plants, and even animals. Biosynthetic gene clusters consist of several genes that encode enzymes, regulatory proteins, transporters, and other factors that are involved in the biosynthesis and regulation of the specialized metabolites. The genes in a BGC are often arranged in a specific order and are regulated by a common set of control elements (Medema *et al.*, 2015).

BGC are of tremendous interest in biotechnology and drug development, as they represent a potential source of new molecules with therapeutic applications. By finding and characterising BGCs, scientists can obtain insight into the biosynthesis of natural products and develop strategies for the production of these molecules in the laboratory. Additionally, knowing the regulation of BGCs can lead to the identification of novel techniques to manipulate gene expression and metabolic pathways to manufacture more efficient and effective medications.

Recent advances in genome sequencing and bioinformatics have enabled the identification and characterization of biosynthetic gene clusters in a wide range of organisms, leading to the discovery of numerous new bioactive compounds. This has significant implications for drug discovery and development, as biosynthetic gene clusters represent a rich source of natural products that can be used as leads for the development of new therapeutics (Chen *et al.*, 2020).

Biosynthetic gene clusters (BGCs) encode for the biosynthesis of a complex range of natural products, which can be classified into several different structural classes based on their chemical compositions includes non-ribosomal peptide synthetases (NRPS), polyketide synthases (PKS), terpenes, and ribosomally synthesized and post-translationally modified peptides (RiPPs) (Medema *et al.*, 2015).

NRPS: Non-ribosomal peptide synthetases (NRPSs) are a class of enzymes that are involved in the biosynthesis of non-ribosomal peptides (NRPs) in microorganisms. NRPs are a diverse class of natural products that have a wide range of biological activities, including antimicrobial, antitumor, and immunosuppressive properties. NRPSs consist of a series of modules, each containing several domains that are responsible for specific biochemical transformations. Each module can incorporate one or more amino acid building blocks into the growing peptide chain, which is then modified and cleaved by other domains in the NRPS enzyme. The final product is typically released from the enzyme by a terminal domain or by proteolytic cleavage.

NRPSs are popular targets for natural product discovery as they are known to synthesize a diversity of antibiotics and immunosuppressants with enormous pharmaceutical potential. Understanding the biosynthetic mechanisms and diversity of NRPSs can provide insights into the discovery and engineering of novel peptide natural products for various applications (Duban *et al.*, 2022).

Non-Ribosomal Peptide	Bioactivity	Microorganism	Reference
Cyclosporin A	Immunosuppressant	<i>Tolypocladium inflatum</i>	(Yang <i>et al.</i> , 2018)
Daptomycin	Antibiotic	<i>Streptomyces roseosporus</i>	(Miao <i>et al.</i> , 2005)
Echinomycin	Antitumor	<i>Streptomyces</i>	(LATHAN & VON HOFF, 1984)
Gramicidin S	Antibiotic	<i>Bacillus brevis</i>	(Yang & Yousef, 2018)

Valinomycin	Antibacterial, antifungal, antitumor, and ionophoric activities	<i>Streptomyces</i>	(Matter <i>et al.</i> , 2009)
Bacitracin	Antibiotic	<i>Bacillus subtilis</i>	(Zhu <i>et al.</i> , 2023)
Actinomycin D	Anticancer	<i>Streptomyces parvulus</i>	(Shetty <i>et al.</i> , 2014)

Table 2.1 NRP and its bioactivity

PKS: Polyketide synthases (PKS) are multi-domain enzymes or enzyme complexes that consist of acyltransferase, ketosynthase, and acyl carrier protein domains and are responsible for the biosynthesis of a large class of natural products known as polyketides. PKSs can be classified based on the types of building blocks they incorporate and the number and organization of their domains and modules. There are three types of PKSs: type I, type II, and type III (Wang *et al.*, 2020).

Type I PKSs are the largest and most complex of the three classes, consisting of multiple modules that function in a colinear manner to synthesize polyketides with a high degree of structural complexity and diversity. Type II PKSs are smaller and simpler than Type I PKSs and consist of discrete enzymes that each catalyse a single biosynthetic step in the production of the polyketide product. Type III PKSs are the simplest of the three classes and consist of a single enzyme that catalyses the entire biosynthetic pathway to produce a polyketide product.

The biosynthetic diversity of PKSs makes them attractive targets for drug discovery and biotechnology applications. The structural complexity and diversity of polyketides produced by PKSs make them valuable as potential leads for the development of new drugs, and the engineering of PKSs has the potential to produce novel polyketides with improved therapeutic properties (Risidian *et al.*, 2019).

Polyketide	Bioactivity	Microorganism	Reference
Erythromycin	Antibiotic	<i>Saccharopolyspora erythraea</i> (formerly <i>Streptomyces erythraeus</i>)	(Cortes <i>et al.</i> , 1990)
Clarithromycin	Antibiotic	Semi-synthetic derivative of erythromycin	(Park <i>et al.</i> , 2010)
Azithromycin	Antibiotic	Semi-synthetic derivative of erythromycin	
Tacrolimus	Immunosuppressant	<i>Streptomyces tsukubaensis</i>	(Ordóñez-Robles <i>et al.</i> , 2018)
Amphotericin	Antifungal agent	<i>Streptomyces nodosus</i>	(Huang <i>et al.</i> , 2021)
Epothilone	Anticancer agent	<i>Sorangium cellulosum</i>	(Forli, 2014)
Lovastatin	Cholesterol-lowering drug	<i>Aspergillus terreus</i>	(Yin <i>et al.</i> , 2016)
Rapamycin	Immunosuppressant	<i>Streptomyces hygroscopicus</i>	(Park <i>et al.</i> , 2010)

Table 2.2 Polyketide and its Bioactivity

Terpene: Terpenes are a large and diverse class of organic compounds found in many natural sources, including plants, animals, and microorganisms and they are synthesized by terpene synthase (TPS) enzyme. They are derived from isoprene units, which are five-carbon building blocks that can be joined together in different ways to form a wide variety of structures. They also have a range of biological activities, including antibacterial, antifungal, and anti-inflammatory effects. Because of these properties, terpenes have been studied for their potential therapeutic applications in medicine.

RiPPs: Ribosomally synthesized and post-translationally modified peptides (RiPPs) are a large class of secondary metabolites that have garnered scientific attention. They are synthesized by ribosomes and undergo post-translational modifications to produce a diverse range of structures and functions. RiPPs can have a wide range of biological activities, including antibiotic, antifungal, and antitumor activities. Genome mining strategies have been developed to identify RiPP BGCs, including RiPP-PRISM, RiPPER, and RiPPMiner-Genome. These tools integrate into the identification of BGCs and predict structures for various families of RiPPs (Russell & Truman, 2020).

RiPPs are classified into various families based on their structural features, including lasso peptides, lanthipeptides, thiopeptides, bottromycins, and cyanobactins.

RiPP Compound	RiPP Class	Bioactivity	Microorganisms
Nisin	Lanthipeptide	Antimicrobial activity	<i>Lactococcus lactis</i>
Thiostrepton	Thiopeptide	Antibacterial activity	<i>Streptomyces azureus</i>
Lacticin 481	Lanthipeptide	Antibacterial	<i>Lactococcus lactis</i>
Microcystin	Cyanobactin	Cytotoxic and hepatotoxic activity	Cyanobacteria
Microcin J25	Lasso peptide	Antibacterial	<i>Escherichia coli</i>
Patellamide A	Cyanobactin	Antitumor activity	<i>Prochloron</i> sp.

Table 2.3 RiPP and its bioactivity

Other BGCs: There are many other types of BGCs that encode for the biosynthesis of diverse natural products, including alkaloids, siderophores, and bacteriocins, Saccharide, RiPP-like, Nucleotide, Phosphonate, Betalactone, Cyanobactin, Polyphenol, Glycopeptide, Lipopeptide, Alkaloid, Butyrolactone etc.

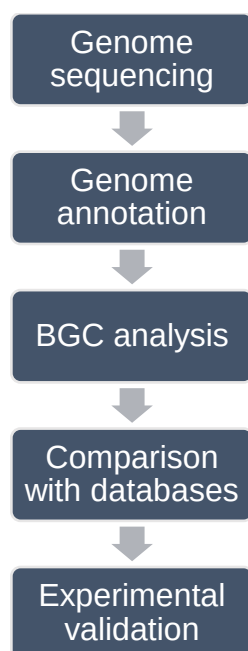
These different classes of BGCs often share common features, such as the presence of large, multimodular enzymes that catalyse the biosynthesis of complex natural products. However, they also have unique characteristics that reflect the specific chemical structures and biological activities of the natural products they produce.

2.5 Genomic Approaches

Genome-based methods have become a powerful tool for uncovering new bioactive compounds. One of the most effective methods in this field is *in silico* genome mining, which involves the computational analysis of genomic data to identify gene clusters that are responsible for the production of natural products (Albarano *et al.*, 2020).

The key steps involved in *in silico* genome mining:

The first step in *In silico* genome mining is to obtain the genome sequence of the halophilic bacterium of interest. This can be done using various sequencing technologies, such as Illumina, PacBio, or Oxford Nanopore. Once the genome sequence is obtained, the next step is to annotate the genome to identify genes and gene clusters that are involved in biosynthesis. This can be done using tools such as RAST, Prokka, or NCBI's Prokaryotic Genome Annotation Pipeline. The annotated genome is then analyzed to identify BGC using specialized software programs. Some commonly used tools for BGC analysis include antiSMASH, PRISM, or MIBiG. These programs search for specific genes and motifs that are associated with the biosynthesis of various natural products, such as polyketides, nonribosomal peptides, or terpenes. The identified BGC can be further analysed by comparing them to existing databases of natural product biosynthetic gene clusters. This can help to determine the structural and functional diversity of the BGC, as well as their potential for producing novel compounds. Finally, the predicted BGC can be experimentally validated by expressing the genes in a heterologous host, such as *E. coli* or *Streptomyces*, and analysing the resulting metabolites using various analytical techniques, such as nuclear magnetic resonance spectroscopy and mass spectrometry (Alam *et al.*, 2022).



Compared to conventional methods of natural product discovery, this strategy has a number of advantages, including the capacity to access the biosynthetic potential of microorganisms that cannot be cultivated and the ability to quickly identify novel compounds with desirable characteristics.

Traditional Natural product Discovery	<i>In silico</i> Genome Mining of BGCs
Laborious and time-consuming	Rapid and high-throughput
The ability to cultivate microorganisms has limitations.	utilises the biosynthetic potential of unculturable microorganisms
Often results in the rediscovery of known compounds	discovers novel compounds with the desired properties
Large amounts of sample materials are required	Only genomic data is required

Table 2.4 Comparison of traditional natural product discovery and *In silico* genome mining of BGCs.

Advantages of *In silico* genome mining:

High throughput	Increased specificity	Access to untapped resources	Reducing the environmental impact
• Genome-based techniques are capable of rapidly analyse vast amount of genomic data, enabling high-throughput screening of promising drug candidates.	• Gene clusters specific to a certain class of natural products can be identified through in silico genome mining, which can increase the specificity of drug discovery efforts.	• Genome-based approaches give researchers access to untapped natural product resources that might not have been identified using conventional techniques.	• The conventional method of isolating natural products from environmental samples can have a significant environmental impact, due to the use of large amounts of solvents, energy-intensive laboratory equipment, and the generation of hazardous waste. On the other hand, genome-based methods only need a small quantity of genetic material.

However, *in silico* genome mining also has several limitations. Firstly, it relies heavily on the quality and completeness of the genomic data available, and incomplete or low-quality genomes can lead to inaccurate predictions. Secondly, it can be difficult to distinguish between active and inactive BGCs, and further experimental validation is necessary to confirm the activity of predicted BGCs. Despite these limitations, there have been several successful discoveries of novel bioactive molecules using *in silico* genome mining (Scherlach & Hertweck, 2021).

2.6 Tools and Databases

antiSMASH	https://antismash.secondarymetabolites.org	(Blin <i>et al.</i> , 2019)
Artemis	https://www.sanger.ac.uk/tool/artemis/	(Carver <i>et al.</i> , 2011)
ClusterFinder	https://github.com/petercim/ClusterFinder	(Cimermancic <i>et al.</i> , 2015)
NP.searcher	https://dna.sherman.lsi.umich.edu/	(Li <i>et al.</i> , 2009)
BAGEL4	http://bagel4.molgenrug.nl/	(van Heel <i>et al.</i> , 2018)

CLUSEAN	https://bitbucket.org/tilmweber/clusean/src/master/	(Blin <i>et al.</i> , 2017)
EvoMining	https://github.com/nselem/EvoMining/wiki	(Sélem-Mojica <i>et al.</i> , 2019)
ARTS	https://arts.ziemertlab.com	(Mungan <i>et al.</i> , 2020)
RiPPMiner	http://www.nii.ac.in/rippminer.html	(Agrawal <i>et al.</i> , 2017)
SBSPKS	http://www.nii.ac.in/sbspks2.html	(Anand <i>et al.</i> , 2010)
PRISM	https://github.com/magarveylab/prism-releases	(Skinnider <i>et al.</i> , 2020)
BiG-SCAPE	https://github.com/medema-group/BiG-SCAPE	(Navarro-Muñoz <i>et al.</i> , 2019)
MIBiG	https://mibig.secondarymetabolites.org/	(Terlouw <i>et al.</i> , 2022)
BiG-SLICE	https://github.com/medema-group/bigslice	(Kautsar <i>et al.</i> , 2021)
IMG-ABC	https://img.jgi.doe.gov/abc	(Palaniappan <i>et al.</i> , 2019)
DoBISCUIT	http://www.bio.nite.go.jp/pks/	(Ichikawa <i>et al.</i> , 2012)
Bactibase	http://bactibase.pfba-lab-tun.org	(Hammami <i>et al.</i> , 2010)
NaPDoS	https://npdomainseeker.sdsc.edu/	(Ziemert <i>et al.</i> , 2012)
DeepBGC	https://github.com/Merck/deepbgc	(Hannigan <i>et al.</i> , 2019)

Table 2.5 Tools and Databases

2.6.1 Description

antiSMASH - Antibiotics & Secondary Metabolites Analysis Shell (antiSMASH) is a web-based tool that uses a combination of algorithms and rules to identify and annotate BGCs in microbial genomes. It can predict the biosynthetic pathways of various types of natural products, including polyketides, nonribosomal peptides, and terpenes. The tool generates detailed reports that include predicted gene clusters,

product structures, and biosynthetic pathway diagrams. It is widely used by researchers and has been cited in hundreds of scientific publications (Blin *et al.*, 2019).

Artemis - It is a genome browser and annotation tool that can be used to visualize and annotate bacterial and archaeal genomes. It allows users to display genomic features such as genes, regulatory elements, and biosynthetic gene clusters, as well as to add their own annotations and notes to the genome. It also provides tools for searching, sorting, and filtering the genome data, as well as for comparing and analyzing multiple genomes (Carver *et al.*, 2011).

MIBiG - The Minimum Information about a Biosynthetic Gene cluster (MIBiG) is a public database that contains annotated BGCs from bacterial and fungal genomes. It provides detailed information about each BGC, including the biosynthetic gene cluster family, predicted product, and chemical structure. It also includes links to other databases, such as NCBI and UniProt, for additional information about the genes and proteins involved in biosynthesis. MIBiG is an important resource for researchers working in natural product discovery and development (Terlouw *et al.*, 2022).

ClusterFinder - It is a software tool for the identification of BGCs in uncharacterized genomic sequences. The tool uses a comparative genomics approach to identify conserved gene clusters that are likely to be involved in biosynthesis. It is based on the assumption that biosynthetic gene clusters are often located in genomic regions that are conserved across different organisms. ClusterFinder generates reports that include predicted gene clusters, product structures, and biosynthetic pathway diagrams. The tool can be used to identify new BGCs in microbial genomes (Cimermancic *et al.*, 2015).

CLUSEAN - CLuster SEquence Analyzer (CLUSEAN) is a web-based tool for the identification and annotation of biosynthetic gene clusters in bacterial and fungal genomes. It uses a combination of genome mining algorithms to identify biosynthetic gene clusters, including antiSMASH, ClusterFinder, and SMURF. It also includes tools for predicting the chemical structures of the compounds produced by the clusters, as well as for comparing and analyzing multiple clusters (Weber *et al.*, 2009).

BiG-SCAPE - Biosynthetic Gene Similarity Clustering and Prospecting Engine (BiG-SCAPE) is a web-based tool for the identification and clustering of biosynthetic gene clusters in bacterial and fungal genomes. It uses a combination of sequence similarity searches and network clustering algorithms to group related biosynthetic gene clusters into families, and provides a range of analytical tools for exploring the predicted clusters (Navarro-Muñoz *et al.*, 2019).

RiPPMiner - This is a tool for the prediction of ribosomally synthesized and post-translationally modified peptide (RiPP) gene clusters in bacterial genomes. RiPPMiner uses a machine learning approach to identify putative RiPP gene clusters based on conserved genomic features. The tool can predict the substrate specificities of individual RiPP biosynthetic enzymes and can generate biosynthetic pathway models (Agrawal *et al.*, 2017).

BAGEL4 - This is a tool for the identification of biosynthetic gene clusters in Gram-positive bacterial genomes that produce bacteriocins, which are antimicrobial peptides. BAGEL4 uses a combination of sequence similarity searches and machine learning algorithms to predict the locations of putative bacteriocin biosynthetic gene clusters. The tool can predict the structures of the corresponding bacteriocins and can generate biosynthetic pathway models (van Heel *et al.*, 2018).

ClustScan - It is a tool for the identification of biosynthetic gene clusters in bacterial genomes based on conserved genomic features. ClustScan uses a combination of sequence similarity searches and machine learning algorithms to predict the locations of putative biosynthetic gene clusters. The tool can be used to predict the biosynthesis of a wide range of natural products, including polyketides, nonribosomal peptides, terpenes, and others (Starcevic *et al.*, 2008).

NP.searcher - This is a web-based tool for the identification and annotation of biosynthetic gene clusters in bacterial and fungal genomes. NP.searcher uses a combination of sequence similarity searches and comparative genomics to predict the biosynthesis of a wide range of natural products, including polyketides, nonribosomal peptides, and others. The tool provides detailed annotations of the predicted biosynthetic gene clusters, including predicted chemical structures, biosynthetic pathways, and potential gene functions (Li *et al.*, 2009).

DeepBGC - It is a machine learning-based tool for the prediction of biosynthetic gene clusters in bacterial genomes. DeepBGC uses a deep neural network to predict the locations of putative biosynthetic gene clusters and to classify them based on the type of natural product they are likely to produce. The tool can predict the structures of the corresponding natural products and can generate biosynthetic pathway models (Hannigan *et al.*, 2019).

SBSPKS - Self-Organizing Map-Based Polyketide Synthase Prediction System is a web-based tool for the prediction of polyketide synthase (PKS) biosynthetic gene clusters in bacterial and fungal genomes. SBSPKS uses a self-organizing map algorithm to predict the locations and types of PKS biosynthetic gene clusters, and provides a range of analytical tools for exploring the predicted clusters (Anand *et al.*, 2010).

NaPDoS - Natural Product Domain Seeker (NaPDoS) is a web-based tool for the identification and annotation of natural product biosynthetic gene clusters in bacterial and fungal genomes. It uses a combination of sequence similarity searches, domain analysis, and other algorithms to predict the locations and types of biosynthetic gene clusters, and provides a range of analytical tools for exploring the predicted clusters (Ziemert *et al.*, 2012).

EvoMining - It is a web-based tool for the prediction and analysis of biosynthetic gene clusters in bacterial and fungal genomes. EvoMining uses a combination of sequence similarity searches, genomic context analysis, and phylogenetic analysis to predict the locations and types of biosynthetic gene clusters, and provides a range of analytical tools for exploring the predicted clusters (Sélem-Mojica *et al.*, 2019).

Bactibase - It is a specialized database that focuses on the identification and annotation of bacteriocin biosynthetic gene clusters in bacterial genomes. It contains information on over 2,000 bacteriocin gene clusters from a wide range of bacterial species. Each gene cluster is annotated with information on the gene content, protein sequences, and functional domains. The database also provides tools for searching and visualizing the data, as well as for predicting bacteriocin biosynthetic gene clusters in new bacterial genomes (Hammami *et al.*, 2010).

IMG-ABC - Integrated Microbial Genomes Atlas of Biosynthetic gene Clusters (IMG-ABC) is a database and analysis platform for biosynthetic gene clusters in microbial genomes. It provides a comprehensive collection of biosynthetic gene clusters from over 45,000 microbial genomes. The database includes tools for browsing, searching, and analyzing the data, as well as for comparing and visualizing the gene clusters across multiple genomes. It also includes tools for predicting the secondary metabolites that are produced by the clusters, based on their biosynthetic pathways (Palaniappan *et al.*, 2019).

DoBISCUIT - Database of BloSynthesis clusters CUrated and InTegrated (DoBISCUIT) is a database and web-based tool for the prediction and analysis of biosynthetic gene clusters in bacterial and fungal genomes. It includes a comprehensive collection of biosynthetic gene clusters from over 1,500 bacterial and fungal genomes, along with annotations and functional predictions for each cluster. The database can be searched and filtered based on various criteria, such as gene content, biosynthetic pathway, and taxonomic classification. It also includes tools for predicting biosynthetic gene clusters in user-submitted genomes, as well as for visualizing and comparing the gene clusters across multiple genomes (Ichikawa *et al.*, 2012).

CHAPTER 3

AIM AND SCOPE

3.1 Aim

The aim of this project is to use *in silico* genome mining to identify biosynthetic gene clusters in halophilic bacteria and explore their diversity and distribution. By analysing publicly available genomic data, we aim to identify potential novel biosynthetic gene clusters that could be involved in the production of bioactive compounds. This project is significant because halophilic bacteria have adapted to extreme environments, and they are known to produce bioactive compounds that have potential for biotechnological and medicinal applications. Therefore, this project has the potential to uncover new and exciting bioactive compounds that could revolutionize the field of drug discovery and biotechnology.

3.2 Scope

The scope of this project is to investigate the potential biotechnological uses and medicinal applications of halophilic bacteria and their biosynthetic gene clusters. We will focus on identifying and characterizing biosynthetic gene clusters in halophilic bacteria, analysing their diversity and distribution, and exploring the potential bioactivities of the compounds they may produce. The project will involve the use of bioinformatics tools and techniques, as well as literature review and data analysis to achieve its objectives. We will also collaborate with experts in the field of natural products chemistry to validate the findings of our *in silico* analysis and identify any potential bioactive compounds. The outcomes of this project have the potential to contribute to the discovery of new natural products with biotechnological and medicinal potential, and could lead to the development of new drugs and biotechnological products.

CHAPTER 4

MATERIALS AND METHODS USED

4.1 Data Collection Using HaloDom

To collect the data for this study, we accessed the HaloDom website. HaloDom is a database of halophilic organisms across all life domains that provides taxonomy information, halotolerance classification, salinity, and genomic information via NCBI and other links. It was created in June 2017. The database contains 1267 entries of halophiles across all life domains. There are 282 Archaea, 669 Bacteria and 314 Eukarya currently registered in the database. HaloDom expands considerably compared to previous halophile databases in terms of coverage, including representatives from all life domains. This makes it a more comprehensive resource for researchers studying halophiles and their adaptations to high salt concentrations. The database is regularly updated with new information, making it a valuable resource for researchers studying halophiles (Loukas *et al.*, 2018).

4.1.1 Importing Data into Ms Excel

We aimed to identify and characterize biosynthetic gene clusters from halophilic bacteria using a bioinformatics approach. To do this, we accessed the HaloDom database and imported all available data into Microsoft Excel. We then sorted and arranged the data to extract only information pertaining to halophilic bacteria.

After obtaining a total of 669 bacterial entries, we further checked for their genome availability using the NCBI browser. We found that out of the 669 entries, only 454 had complete, contig, or scaffold genomes. We included all three types of genomes in our study to maximize the diversity of our dataset and ensure that we were able to capture a broad range of biosynthetic gene clusters from halophilic bacteria.

This project represents an exploratory study aimed at identifying and characterizing biosynthetic gene clusters from halophilic bacteria using a bioinformatics approach. By utilizing the Halodom database and NCBI browser, we were able to assemble a dataset of 454 halophilic bacteria with available genomes, which we will analyze

using a variety of bioinformatics tools and techniques to identify novel biosynthetic gene clusters with potential biotechnological applications.

4.1.2 Obtaining Genomic Information

The genomic information for 454 entries were obtained from NCBI in GenBank Flat File Format (.gbff), we saved the downloaded files to a local directory for further analysis. The GenBank format provided a comprehensive annotation of the genome, including the nucleotide sequences of genes, their functions, and their positions in the genome.

Before proceeding with the analysis of the downloaded genomes, we conducted a screening process to remove any duplicates from the dataset. This was an important step to ensure that our analysis was not biased and that we were working with a clean and reliable dataset.

The screening process involved comparing the genome IDs of each entry in the dataset and removing any entries with duplicate IDs. This resulted in a final dataset of non-redundant halophilic bacterial genomes with comprehensive genomic annotations, which we used to identify and characterize biosynthetic gene clusters using a bioinformatics approach.

By screening for duplicates, we were able to avoid potential errors and biases that could have arisen from including redundant genomic data in our analysis. This ensured that our findings are accurate and reliable, and can be used to inform future research on halophilic bacteria and their potential applications.

4.2 Identification of BGCs

In our analysis of the non-redundant halophilic bacterial genomes, we utilized antiSMASH, a bioinformatics tool, to identify biosynthetic gene clusters (BGCs) in the downloaded genomes.

4.2.1 antiSMASH

antiSMASH is an open-source tool that allows the rapid genome-wide identification, annotation, and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes. It is powered by several open source tools: NCBI BLAST+, HMMer 3, Muscle 3, Glimmer 3, FastTree, TreeGraph 2, Indigo-depict,

PySVG and JQuery SVG. It integrates and cross-links with a large number of *in silico* secondary metabolite analysis tools that have been published earlier. It was developed as a collaboration between the Department of Microbial Physiology and Groningen Bioinformatics Centre of the University of Groningen, the Department of Microbiology of the University of Tübingen, and the Department of Bioengineering and Therapeutic Sciences at the University of California, San Francisco. The latest version of antiSMASH 6.0, improves cluster detection and comparison, making it an even more powerful tool for researchers studying secondary metabolite biosynthesis gene clusters. antiSMASH can be run online or installed locally on a computer. Secondary metabolites are bioactive compounds produced by microorganisms that have potential applications in various fields, including medicine and biotechnology.

4.2.2 antiSMASH Input Data

The ideal input for antiSMASH is an annotated nucleotide file in Genbank format or EMBL format. Users can either upload a local GenBank/EMBL file manually or enter the GenBank/RefSeq accession number of the sequence for antiSMASH to upload it.

If no gene annotations are available for the sequence, antiSMASH recommends running the sequence through an annotation pipeline like RAST to obtain GBK/EMBL files with high-quality annotations. Alternatively, users can provide a FASTA file containing a single sequence, and antiSMASH will generate a preliminary annotation using Prodigal and use that to run the rest of the analysis. Users can also provide gene annotations in GFF3 format.

If users are creating their GBK/EMBL/FASTA file manually, they should do so in a plain text editor like Notepad or Emacs and save their files as "All files (.)", ending with the correct extension (for example ".fasta", ".gbk", or ".embl". Overall, antiSMASH can work with different formats of data, but it is essential to provide properly formatted input files to ensure accurate and reliable results.

4.2.3 Extra features of antiSMASH

antiSMASH has several extra features that can be toggled on or of

KnownClusterBlast analysis: It is an algorithm used to compare identified biosynthetic gene clusters (BGCs) against the manually curated set of known BGCs from the MIBiG (Minimum Information about a Biosynthetic Gene cluster) repository. The MIBiG is a hand-curated data collection of biosynthetic gene clusters that have been experimentally characterized. The KnownClusterBlast analysis helps to identify known BGCs that are similar to the identified BGCs, which can provide insights into the biosynthetic pathways and chemical structures of the secondary metabolites (Blin *et al.*, 2018).

Subcluster Blast analysis: This feature is used to identify conserved operons with known functions within identified biosynthetic gene clusters (BGCs). It helps to identify subclusters within the identified BGCs that are similar to known subclusters with known functions in the MIBiG (Minimum Information about a Biosynthetic Gene cluster) repository (Blin *et al.*, 2017).

ClusterBlast analysis: The identified clusters are searched against a comprehensive gene cluster database, and similar clusters are identified using the MultiGeneBlast algorithm. The MultiGeneBlast algorithm runs BlastP using each amino acid sequence from a detected gene cluster as a query on a large database of predicted protein sequences from secondary metabolite biosynthetic gene clusters, and pools the results to identify the gene clusters that are most homologous to the gene cluster that was detected in the query nucleotide sequence (Blin *et al.*, 2013).

Cluster Pfam analysis: It is used to analyze each gene product encoded in the detected biosynthetic gene clusters (BGCs) against the PFAM database. The PFAM database is a collection of protein families that are used to annotate each gene in the BGCs. The PFAM analysis is often used in combination with other features such as ClusterBlast and Pfam-based GO term annotation to provide a more comprehensive view of the BGCs.

Pfam-based GO term annotation: It is used to predict the functions of the proteins encoded by the genes in the BGCs and to gain insights into the biosynthetic pathways and chemical structures of the secondary metabolites encoded by the BGCs.

MIBiG cluster comparison: It is used to compare the identified BGCs against the known BGCs in the MIBiG database. The comparison is done using Pfam composition similarity, and distances are calculated with an adapted Jaccard and domain duplication.

ActiveSiteFinder: This module is used to identify and annotate conserved enzyme active sites in the protein structures. It is an automated active site detection, docking, and scoring protocol for proteins with known structures.

TIGRFam: It is a manually curated protein family database that can be used to scan regions in antiSMASH.

To use antiSMASH, the downloaded genome files were uploaded to the software as input. The analysis parameters were set to detect BGCs with a relaxed detection strictness, which allows for the identification of well-defined clusters containing all required parts, as well as partial clusters missing one or more functional parts. This helps to ensure that as many BGCs as possible are identified in the input genomes.

4.2.4 *antiSMASH OUTPUT*

The extra features option was also turned on, which allows antiSMASH to detect additional genes that may play a role in the biosynthesis, regulation, or modification of secondary metabolites.

The antiSMASH full output was downloaded as the results on the public web server are only kept for one month and will be deleted afterwards. The result page offers a comprehensive overview of all the output results in a single page, providing detailed information about each biosynthetic gene cluster detected in the input sequences. It also displays the type classification of each detected gene cluster, as well as its location and all the genes present in the cluster.

4.2.4.1 *Manual Curation of Data into Excel*

After obtaining the results of the antiSMASH analysis, we manually curated the data by categorizing and organizing the detected biosynthetic gene clusters in an Excel spreadsheet. This process involved reviewing and sorting through the results to ensure that only the relevant data were included in the analysis. We removed any duplicate entries, corrected any errors or inconsistencies in the data, and standardized the format of the entries to facilitate analysis.

Once the data were curated, we performed statistical analysis to identify trends and patterns in the biosynthetic gene clusters. Specifically, we examined the frequency and distribution of the different types of gene clusters detected in the halophilic bacterial genomes, as well as their functional properties. We used various statistical techniques, such as frequency analysis and regression analysis, to identify significant trends and relationships in the data.

The goal of this step was to gain a deeper understanding of the biosynthetic gene clusters present in the halophilic bacterial genomes and to identify any potential biotechnological applications of these gene clusters. The results of this analysis provide valuable insights into the diversity and functional properties of these gene clusters and can serve as a basis for future studies in this field.

4.3 Phylogenetic tree Construcion

To perform the phylogenetic analysis, we first searched and downloaded the 16S rRNA gene sequences of halophilic bacteria from the NCBI database. The downloaded sequences were then aligned using BioEdit.

4.3.1 BioEdit

BioEdit is a popular and user-friendly software tool for biological sequence alignment and editing, specifically designed to assist researchers in aligning and editing DNA and protein sequences. Its primary purpose is to provide a single platform that can handle most basic sequence and alignment editing and manipulation functions that researchers may need to perform regularly. It also offers basic sequence analysis functions. The software has an intuitive multiple document interface that facilitates easy alignment and manipulation of sequences. It supports various file formats and provides users with several shortcuts for the editor window, enhancing usability. It is available at <https://bioedit.software.informer.com/7.2/>

We used the BioEdit software to manually edit and clean the sequences, removing any ambiguous bases, sequencing errors, or gaps. During the alignment process, the 5' and 3' end regions of some sequences appeared to be empty, and therefore we trimmed these regions to obtain a consistent sequence length across all sequences. After alignment, we imported the sequences into the MEGA software to construct the phylogenetic tree.

4.3.2 MEGA

MEGA (Molecular Evolutionary Genetics Analysis) is a user-friendly bio-computational software for sequence analysis and phylogenetic analysis. It is an integrated tool for conducting automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, and testing evolutionary hypotheses. It is user-friendly and has a graphical user interface that allows viewing and editing of the phylogenetic trees. It also constructs trees using various methods, such as Maximum Likelihood, Neighbor Joining, and UPGMA. It is available for free download on various websites and is licensed as proprietary freeware. It is available at <https://www.megasoftware.net/>

In our study, we chose the neighbour-joining method to construct the phylogenetic tree for our dataset. We also set the bootstrap value to 1000, which allowed us to estimate the reliability of the branches in the tree. The resulting phylogenetic tree provided insight into the evolutionary relationships among the halophilic bacterial species.

4.4 Distribution of BGCs in Halophilic Bacteria

After inferring the evolutionary relationships between halophilic bacterial genomes based on their 16S rRNA gene sequences using BioEdit, we further analyzed the distribution of biosynthetic gene cluster (BGC) classes across these genomes. This allowed us to identify potential trends in the evolution of BGCs in halophilic bacteria and their relationship to the genomes' ability to produce certain types of natural products. To visualize the distribution of BGC classes across the phylogenetic tree of the 454 halophilic bacterial genomes, we used the Interactive Tree Of Life (iTOL).

4.4.1 iTOL

Interactive Tree of Life (iTOL) is an online tool for the display, annotation, and management of phylogenetic and other trees. It provides a large variety of tree layouts, and users have full control over branch colors, widths, and styles. It allows users to organize their trees into workspaces and projects, and access them from any browser. The software provides 19 dataset types, and users have individually adjustable label fonts, sizes, and styles.

iTOL allows users to create high-quality tree figures for their publications, and it provides direct What-You-See-Is-What-You-Get export of what is displayed on the screen. It also allows users to export their trees into various vector or bitmap formats. The software is optimized for working with large datasets, and it provides a Python API and a command-line tool for interacting with the iTOL server. iTOL annotation editor provides access to iTOL directly from Google Sheets, Microsoft Excel, or LibreOffice Calc, allowing users to easily annotate their trees with additional information. It is widely used in molecular biology research, and it is an essential tool for interpreting biological information from vast amounts of raw sequence data.

4.4.2 Visualization and Annotation

We downloaded the phylogenetic tree in Newick format from the MEGA software. The Newick format is a widely used standard for representing phylogenetic trees in a text file format. Next, we uploaded the Newick format tree file to the Interactive Tree of Life (iTOL) online tool, which is a user-friendly web-based application for visualizing and annotating phylogenetic trees.

To map the distribution of BGC classes onto the phylogenetic tree, we created a BGC dataset that contained information about the types of biosynthetic gene clusters identified in each of the 454 halophilic bacterial genomes. We then uploaded this BGC dataset to iTOL and used its annotation features to map the BGC classes onto the phylogenetic tree. This enabled us to visualize the distribution of different BGC classes across the phylogenetic tree and investigate the evolutionary relationship between the genomes and their potential to produce certain types of natural products.

CHAPTER 5

RESULTS AND DISCUSSION

5.1 Results

5.1.1 Halophilic Bacterial Dataset Analysis

In this study, we conducted a systematic investigation of the biosynthetic potential of 454 halophilic bacterial species using their genome sequences. We obtained the data from HaloDom and downloaded the genomic information of the halophilic bacteria available in public database NCBI.

We analysed a dataset consisting of 454 genomes of halophilic bacteria, that includes 64 complete genomes, 203 scaffold genomes, and 167 contig genomes. To maximize the diversity of our dataset and capture a broad range of biosynthetic gene clusters (BGCs) from halophilic bacteria, we included all three types of genomes in our study. We also created a graph to visualize the distribution of genome availability in our dataset. (Fig 5.1)

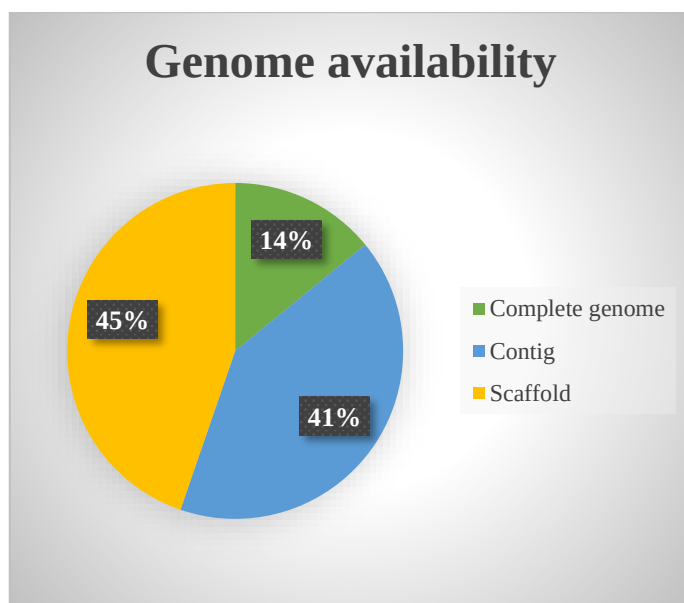


Fig 5.1 Genomic availability of Halophilic bacteria

Based on the halotolerance classification of halophilic bacteria into extreme, moderate, and slight categories, we found that among the 454 genomes, there were 55 extreme halophiles, 350 moderate halophiles, and 49 slight halophiles. The majority of the genomes were of the moderate halotolerance category. (Fig 5.2)

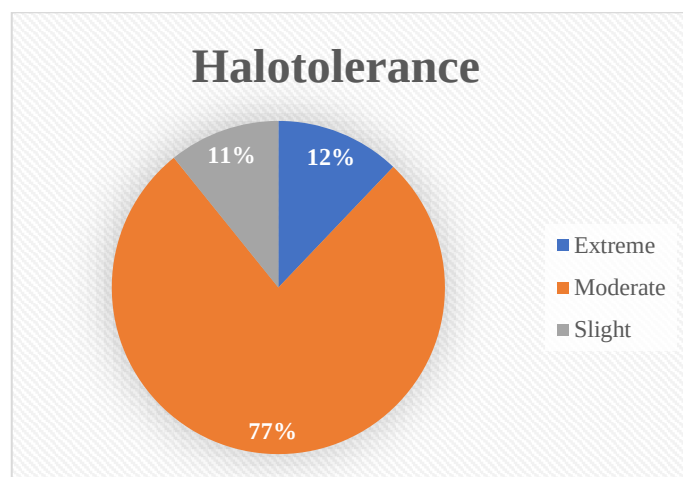


Fig 5.2 Halotolerance classification of halophilic bacteria

We also wanted to understand the distribution of the three halotolerance categories (extreme, moderate, and slight) within our dataset and their genome completeness. To do this, we created a graph that visualized the number of genomes in each halotolerance category with respect to their genome type (complete, scaffold, or contig). (Fig 5.3)

These findings suggest that the completeness of the genomes varied across the halotolerance categories. This information could be useful in future studies to determine the relationship between genome completeness and biosynthetic potential in halophilic bacteria.

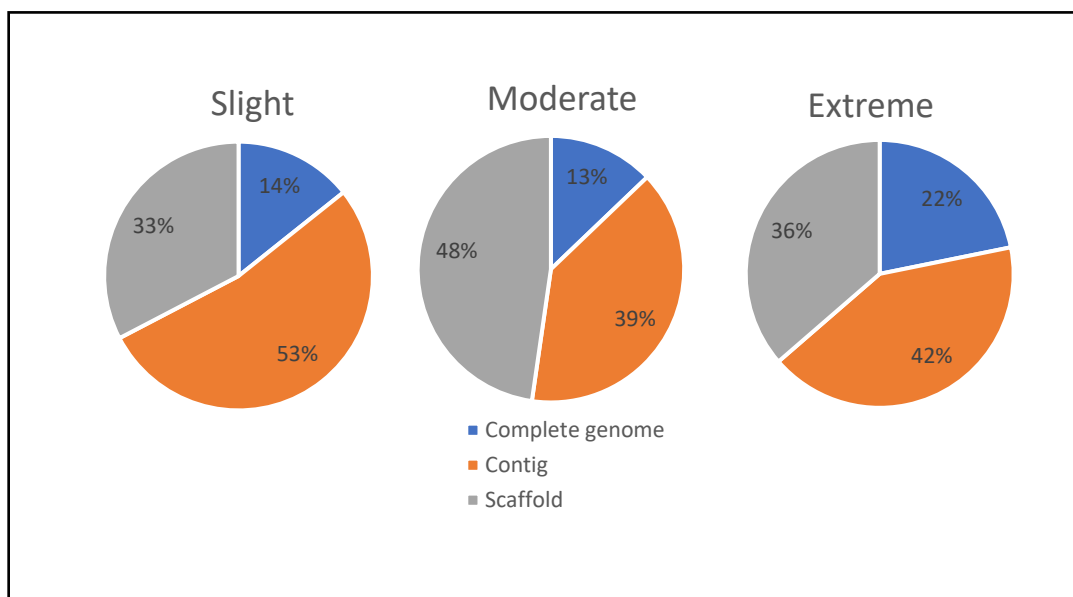


Fig 5.3 Distribution of halotolerance and their genome completeness

5.1.2 Distribution of BGCs in Halophilic Bacteria

Subsequently, we employed the antiSMASH 6.0 bacterial version to extract the biosynthetic gene clusters (BGCs) encoding for natural products from the 454 halophilic bacterial species. Our analysis showed that the halophilic bacteria possessed a wide range of BGCs that have the potential to produce various natural products. To further understand the diversity of these BGCs, we manually curated and categorized them based on their types and functional properties. We performed statistical analyses on the curated BGC dataset to identify trends and patterns in the halophilic bacterial species. These analyses enabled us to determine the most common types of BGCs, their distributions across different halophilic bacterial genomes, and their functional properties.

Our analysis using antiSMASH 6.0 bacterial version revealed a diverse range of biosynthetic gene clusters (BGCs) in the 454 halophilic bacterial species studied. We identified a total of 58 different BGC classes, including acyl_amino_acids, amglyccycl, arylpolyene, betalactone, butyrolactone, CDPS, cyanobactin, cyclic-lactone-autoinducer, ectoine, furan, hserlactone, glycocin, guanidinotides, HgIE-KS, indole, ladderane, lanthipeptide-class-i, lanthipeptide-class-ii, lanthipeptide-class-iii, lanthipeptide-class-iv, lanthipeptide-class-v, LAP, lassopeptide, linaridin, melanin,

microviridin, NAGGN, NAPAA, NRPS, NRPS-like, nucleoside, oligosaccharide, phenazine, phosphonate, PKS-like, proteusin, prodigiosin, ranthipeptide, redox-cofactor, resorcinol, RiPP-like, RRE containing, sactipeptide, siderophore, spliceotide, terpene, thioamitides, thiopeptide, thioamide-NRP, transAT-PKS, transAT-PKS-like, T1PKS, T2PKS, T3PKS, PBDE, PUFA, and tropodithietic-acid. (Fig 5.4)

Interestingly, we also observed variation in the number of BGCs present within each genome, with halophilic bacterial species carrying between 0 and 44 BGCs per genome (mean = 6.11, s.d = 5.59). The highest 20 genomes contained more than 20 BGCs each, while 6 genomes lacked any detectable BGCs. (Table 5.1 and 5.2)

Halophilic bacteria	No. of BGCs	Halotolerance classification	Genome availability
<i>Streptomyces chilikensis</i>	44	Moderate	Scaffold
<i>Nocardiopsis halotolerans</i>	37	Moderate	Contig
<i>Actinopolyspora erythraea</i>	33	Moderate	Complete genome
<i>Haliangium ochraceum</i>	33	Slight	Complete genome
<i>Actinopolyspora biskrensis</i>	32	Extreme	Scaffold
<i>Gynuella sunshinyii</i>	30	Moderate	Complete genome
<i>Actinopolyspora alba</i>	29	Moderate	Scaffold
<i>Nocardiopsis halophila</i>	28	Moderate	Contig
<i>Actinopolyspora mzabensis</i>	27	Extreme	Scaffold
<i>Chlorogloea fritschii</i>	26	Moderate	Contig
<i>Nocardiopsis xinjiangensis</i>	26	Moderate	Contig
<i>Actinopolyspora righensis</i>	24	Extreme	Scaffold
<i>Nocardiopsis terrae</i>	24	Slight	Scaffold
<i>Actinopolyspora mortivallis</i>	23	Moderate	Scaffold
<i>Haloactinospira alba</i>	22	Moderate	Contig
<i>Nocardiopsis kunsanensis</i>	22	Moderate	Contig
<i>Nodularia spumigena</i>	21	Moderate	complete genome
<i>Saccharomonospora paurometabolica</i>	20	Moderate	Scaffold
<i>Saccharomonospora saliphila</i>	20	Moderate	Scaffold
<i>Streptomonospora alba</i>	20	Moderate	Contig

Table 5.1 Top 20 Halophilic bacterial species contains more than 20 BGCs

Halophilic bacteria	No. of BGCs	Halotolerance classification	Genome availability
<i>Desulfovibrio gracilis</i>	0	Moderate	Contig
<i>Mycoplasma marinum</i>	0	Slight	Contig
<i>Mycoplasma todarodis</i>	0	Slight	Contig
<i>Salinicoccus carniancri</i>	0	Moderate	Contig
<i>Salinicoccus halodurans</i>	0	Moderate	complete genome
<i>Salinimonas chungwhensis</i>	0	Moderate	Scaffold
<i>Salinimonas iocasmae</i>	0	Moderate	complete genome

Table 5.2 Seven halophilic bacterial species did not contain BGCs

Our findings revealed that the most common BGCs were ectoine found in 291 genomes, terpene in 268 genomes, betalactone in 171 genomes, T3PKS in 170 genomes, and RiPP-like in 149 genomes. These BGCs were present across a wide range of halophilic bacterial species, highlighting the potential for these organisms to produce these natural products. (Fig 5.4)

We also found that a species may carry multiple copies of the same or different BGC classes. For instance, *Streptomyces chilikensis* was found to carry a diverse array of BGCs including 1 amglyccyl, 4 butyrolactone, 1 indole, 1 melanin, 1 ectoine, 5 NRPS, 4 NRPS-like, 3 RiPP-like, 2 siderophore, 6 terpene, 1 trans-AT-PKS, 13 T1PKS, 1 T3PKS, and 1 other BGCs.

Some BGCs were uncommon and found in only a few genomes. These include CDPS (5 genomes), Lanthipeptide-class-v, Thioamide-NRP (4 genomes), Furan (3 genomes), Cyanobactin, Glycocin, Guanidinotides, Lanthipeptide-class-iv, Nucleoside (2 genomes), Melanin, Microviridin, Spliceotide, PBDE and Tropodithietic acid (1genome).

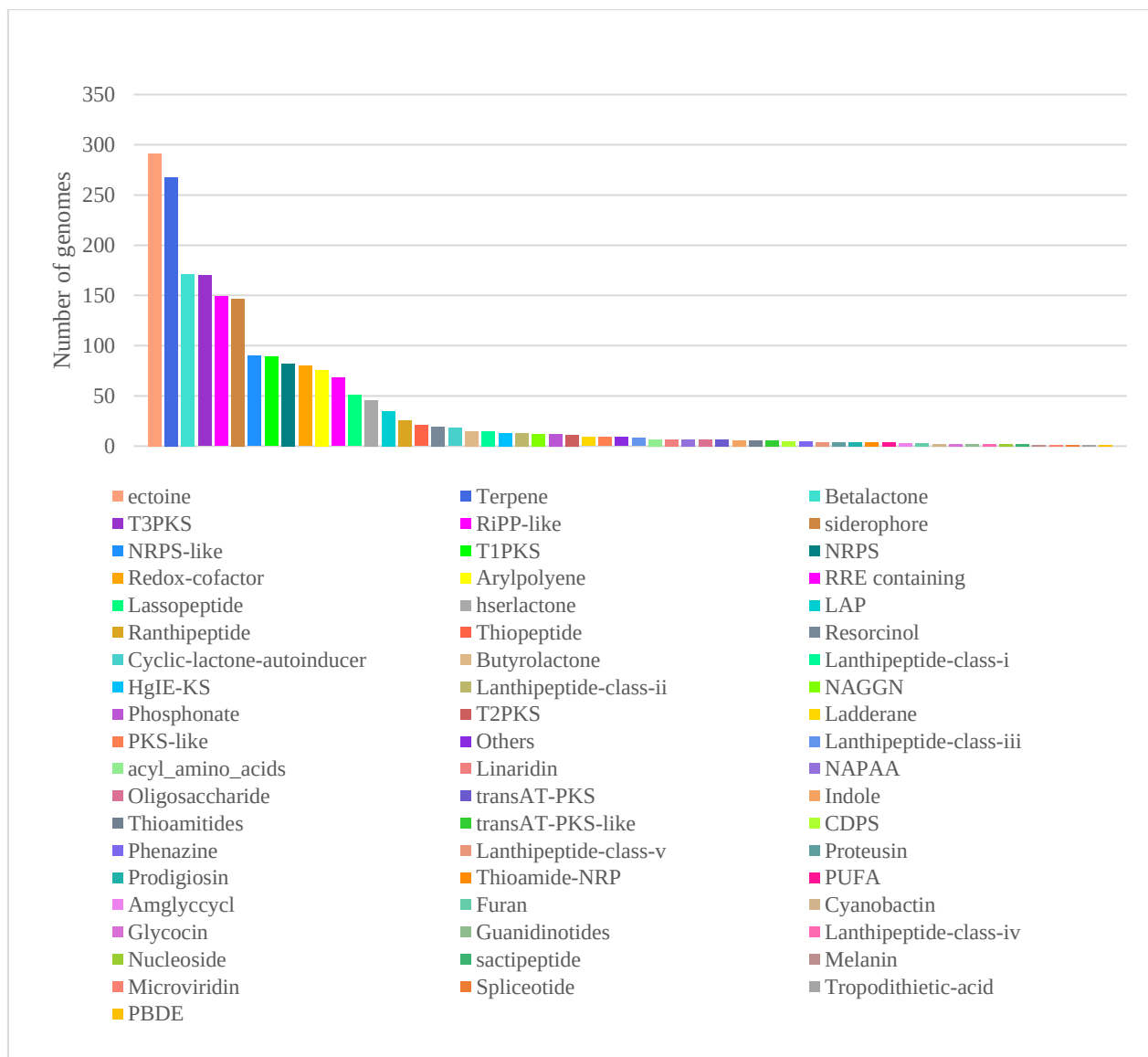


Fig 5.4 Distribution of Biosynthetic Gene clusters (BGCs) in Halophilic bacteria

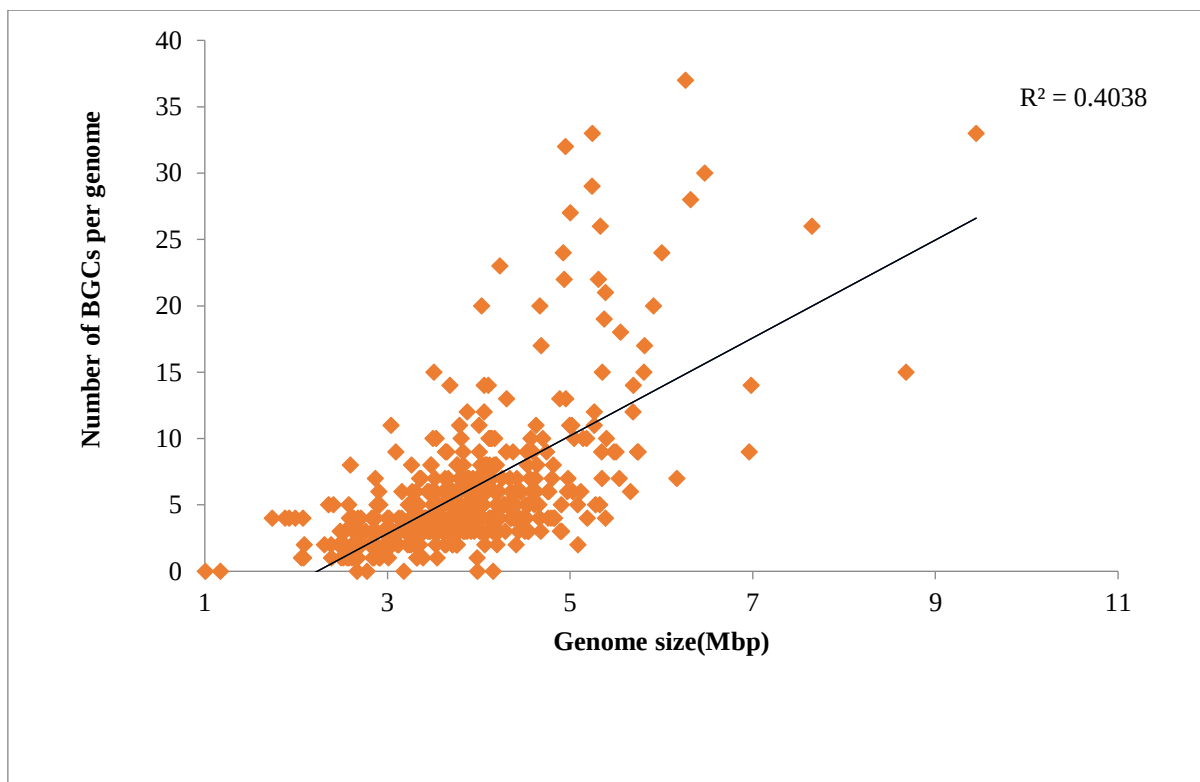


Fig 5.5 Correlation between genome size and BGC

The correlation analysis between genome size and the number of BGCs revealed a moderate positive correlation with an R-squared value of 0.4038. Fig(5.5)

5.1.3 Phylogenetic Diversity of BGCs in Halophilic Bacteria

We utilized both BioEdit and Mega software to construct a phylogenetic tree based on the 16S rRNA of 454 halophilic bacterial species. We used 16S rRNA gene because it is widely used molecular marker for bacterial classification and phylogenetic analysis due to its conserved nature and slow evolution rate.

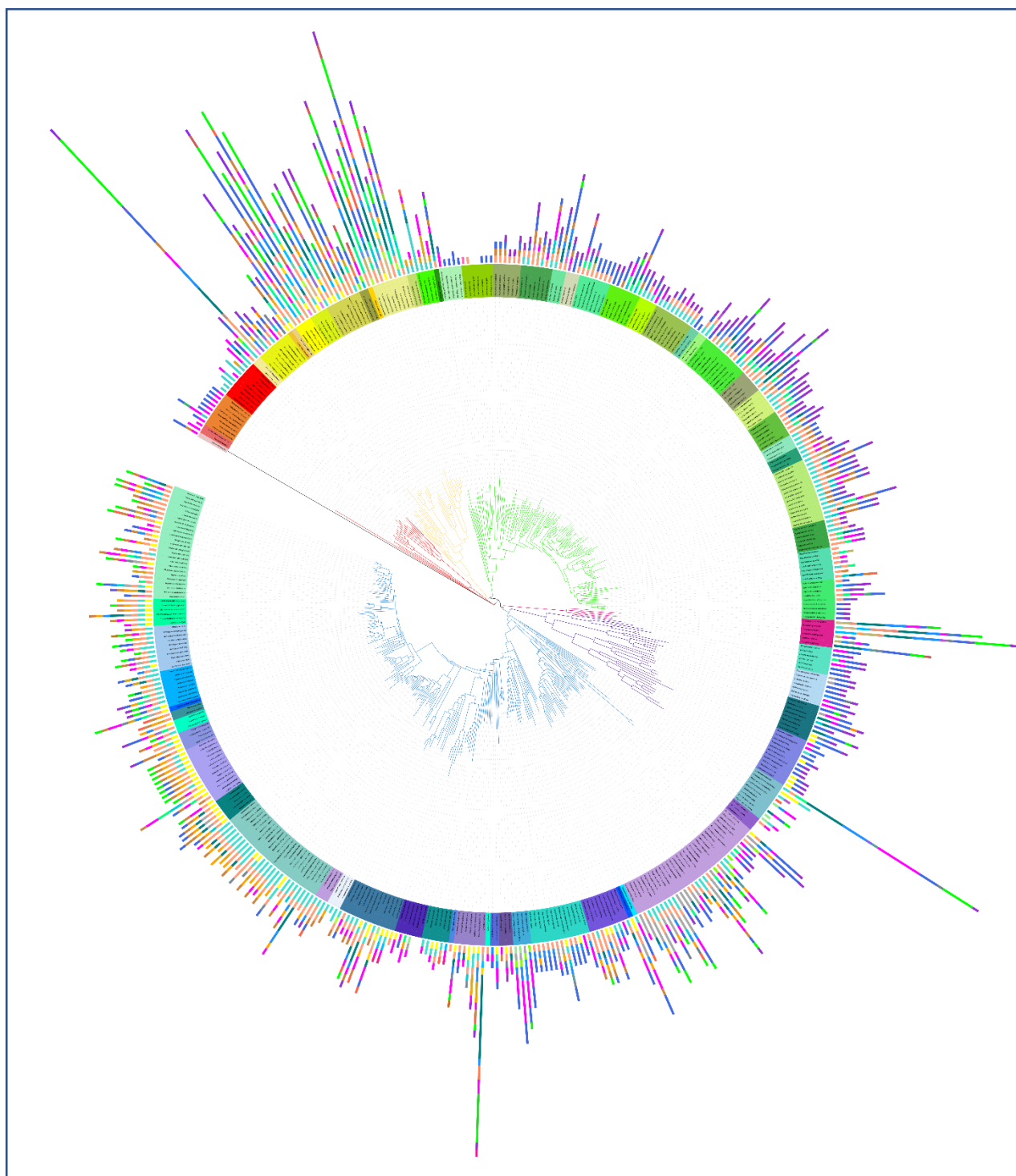


Fig 5.6 Phylogenetic distribution of BGCs in Halophilic bacteria

Biosynthetic gene clusters(BGCs)		
acyl_amino_acids	Lanthipeptide-class-iii	Ranthipeptide
Amglyccycl	Lanthipeptide-class-iv	Redox-cofactor
Arylpolyene	Lanthipeptide-class-v	Resorcinol
Betalactone	LAP	RiPP-like
Butyrolactone	Lasso peptide	RRE containing
CDPS	Linaridin	sactipeptide
Cyanobactin	Melanin	siderophore
Cyclic-lactone-autoinducer	Microviridin	Spliceotide
ectoine	NAGGN	Terpene
Furan	NAPAA	Thioamitides
hserlactone	NRPS	Thiopeptide
Glycocin	NRPS-like	Thioamide-NRP
Guanidinotides	Nucleoside	transAT-PKS
HgIE-KS	Oligosaccharide	transAT-PKS-like
Indole	Phenazine	T1PKS
Ladderane	Phosphonate	T2PKS
Lanthipeptide-class-i	PKS-like	T3PKS
Lanthipeptide-class-ii	Proteusin	Others
	Prodigiosin	PUFA

The phylogenetic tree was constructed using the neighbor joining method and bootstrap resampling with 1000 replicates, based on the alignment of the 16S rRNA gene sequences of the 454 halophilic bacterial genomes. This method was chosen due to its efficiency in handling large datasets and producing reliable results. The resulting tree was visualized using iTOL (Interactive Tree of Life) and analysed to gain insights into the evolutionary relationships and diversity of the halophilic bacterial genomes. The tree showed distinct clusters of related species and

provided a framework for further analysis of the distribution of BGCs among the different lineages.

The phylogenetic tree analysis conducted in this study revealed the presence of six well-defined clades, each of which was represented by a distinct color in the tree. These clades were found to belong to different phyla, highlighting the diversity of bacterial species and their evolutionary relationships. (Fig 5.6)

5.2 DISCUSSION

To begin with, the analysis of our phylogenetic tree revealed six distinct clades that belong to different phyla. These results support our research question of exploring the diversity and distribution of biosynthetic gene clusters (BGCs) in different bacterial phyla. Furthermore, our analysis showed that BGCs are not evenly distributed among the different clades, indicating that certain phyla or clades may have a greater potential for producing bioactive compounds than others.

Clade 1:

Our analysis of clade 1 of the phylum Firmicutes revealed that the 20 species in this clade carry between 0 and 12 different biosynthetic gene clusters (BGCs) per genome (MEAN = 4.1, S.D = 3.5). The most common BGCs found in clade 1 species were RiPP-like, Cyclic lactone-autoinducer, and terpene. However, we also identified several other BGC classes, including Betalactone, RRE containing, NRPS, NRPS-like, Ranthipeptide, sactipeptide, and siderophore.

Our analysis also revealed that the number of BGCs per genome in clade 1 species is relatively low compared to other classes of halophilic bacteria. However, the diversity of BGC classes identified suggests that clade 1 species may still represent a valuable resource for the discovery of novel bioactive compounds.

Clade 2:

We found that Clade 2 consisted of 39 halophilic bacterial species that belonged to the phylum Actinobacteria. These species were found to carry between 3 and 20 different BGCs per genome (mean = 10.51, S.D = 4.547). Among the identified BGCs, Terpene, ectoine, siderophore, T1PKS, T3PKS, NRPS, and RiPP-like were found to be the most common BGCs. The other BGCs found in clade 2 were RRE containing, Lanthipeptide-class-I, Lasso peptide, Butyrolactone, Betalactone, LAP,

T2PKS, Arylpolyene, Others, PKS-like, HgIE-KS, Ladderane, Linaridin, Lanthipeptide-class-iii, NAPAA, Oligosaccharide, Redox-cofactor, Thiopeptide, Indole, Ranthipeptide, CDPS, Phenazine, Amglyccycl, Furan, NAGGN, transAT-PKS, transAT-PKS-like, Guanidinotides, Lanthipeptide-class-ii, Nucleoside, Phosphonate, Resorcinol, Thioamitides, Thioamide-NRP, Hserlactone, Lanthipeptide-class-v, Melanin and Prodigiosin.

Upon further investigation, we found that *Nocardiopsis halotolerans* had the highest number of BGCs, with 20 identified BGCs. This was followed by *Actinopolyspora erythraea* with 19 BGCs, *Nocardiopsis terrae* with 17 BGCs, and *Actinopolyspora mortivallis* with 16 BGCs.

This clade was found to carry a higher number of BGCs compared to other classes of halophilic bacteria.

Our findings suggest that halophilic Actinobacteria may be a particularly rich source of BGCs and bioactive substances, especially in the context of halophilic environments. This observation is consistent with other research that has shown Actinobacteria to produce a broad variety of bioactive compounds with therapeutic potential (Jose *et al.*, 2021). Our analysis provides further support to the concept that Actinobacteria are a promising target for drug discovery and highlights the potential of halophilic environments as a source of novel bioactive compounds.

Nocardiopsis halophila is a halophilic actinomycete that was isolated from soil and has been identified as a potential source of bioactive compounds. Studies have reported that *Nocardiopsis* species produce a variety of bioactive compounds, including antimicrobial agents (Bennur *et al.*, 2015).

A recent study reported the identification and characterization of a fluorinase from *Actinopolyspora mzabensis* (Sooklal *et al.*, 2020).

Clade 3:

Clade 3, which is composed of 139 species belonging to Phylum Firmicutes, exhibited a wide range of BGCs per genome, with the number of BGCs ranging from 0 to 9. The mean and standard deviation for BGCs in this clade were 3.827 and 1.903, respectively. Notably, two genomes in this clade, *Salinicoccus carniancri* and *Salinicoccus halodurans*, contained no BGCs.

The most commonly found BGCs in clade 3 were Terpene, T3PKS, and Ectoine, with 100, 97, and 94 genomes containing these BGCs, respectively. The other BGCs found in clade 3 species were Betalactone, siderophore, RiPP-like, Lasso peptide, NRPS, LAP, RRE containing, Cyclic-lactone-autoinducer, Lanthipeptide-ii, NRPS-like, Ranthipeptide, Phosphonate, T1PKS, Thiopeptide, Glycocin, Lanthipeptide-class-iii, transAT-PKS, hserlactone, lanthipeptide-class-i, lanthipeptide-class-iv, Redox-cofactor, sactipeptide, and transAT-PKS-like.

These results suggest that species in clade 3 have a diverse range of biosynthetic capabilities, with some species producing a relatively small number of BGCs and others producing a much larger number. The presence of a wide variety of BGCs in clade 3 suggests that these species may be a valuable resource for the discovery of novel bioactive compounds. However, further research is needed to investigate the bioactivity of these compounds and their potential for drug discovery.

Clade 4:

Clade 4, consisting of 6 species belonging to the Phylum Cyanobacteria, has been found to carry a considerable number of biosynthetic gene clusters (BGCs) per genome, ranging from 3 to 12. The mean number of BGCs per genome was 8, with a standard deviation of 3.577. The most common BGC found in all species was Terpene. Other BGCs found in clade 4 include NRPS-like, Lanthipeptide-class-ii, NRPS, T1PKS, Lanthipeptide-class-v, Resorcinol, RRE containing, Cyanobactin, HgIE-KS, RiPP-like, Betalactone, Indole, Microviridin, Phosphonate, Proteusin, Spliceotide, transAT-PKS, transAT-PKS-like, T2PKS and T3PKS.

These results suggest that the species in clade 4 may be a valuable source of bioactive compounds for drug discovery, particularly in the context of their ability to produce a diverse array of BGCs. Terpene, being the most common BGC, may have a particularly important role in the biosynthesis of bioactive compounds in these species.

Aphanothece halophytica is a halotolerant alkaliphilic cyanobacterium. Several studies have investigated the bioactivity of this species. One study identified bioactive metabolites from *Aphanothece halophytica* through HPLC and characterized the analytes through ESI-MS and NMR (Vishwakarma, 2013). Another study found that the crude extract of *Aphanothece halophytica* exhibited

bioactivity against Gram-positive bacteria. These findings suggest that *Aphanothece halophytica* has the potential to produce bioactive compounds with antibacterial and cytotoxic properties (Vishwakarma & Rai, 2013).

Further investigation into the bioactivity of these species in clade 4 may lead to the development of new drugs with therapeutic potential.

Clade 5:

Clade 5 was composed of 32 species belonging to three different phyla: Bacteroidetes, Rhodothermaeota, and Balneolaeota. Our analysis reveals that these species carry a relatively low number of biosynthetic gene clusters (BGCs), ranging from 1 to 5 per genome, with a mean of 2.84 and a standard deviation of 0.987. The most common BGCs found in clade 5 species are terpene and T3PKS, which are present in 32 and 26 genomes, respectively. In addition, other BGCs were identified, such as arylpolyene, RiPP-like, RRE-containing, hserlactone, NRPS, NRPS-like, T1PKS, betalactone, LAP, lassopeptide, resorcinol, thioamitides, and thioamide-NRP.

Our findings suggest that although clade 5 species carry a relatively low number of BGCs, they still represent a potential source of bioactive compounds with therapeutic potential. It is worth noting that terpene and T3PKS, the most common BGCs found in these species, have been previously associated with a wide range of biological activities, including antibacterial, antifungal, and anticancer effects.

Clade 6:

Clade 6 consists of 218 species belonging to the phylum Proteobacteria. Results show that Clade 6 species carry between 0 to 14 BGCs per genome, with a mean of 4.61 and a standard deviation of 2.274. The most common BGCs found in Clade 6 are ectoine (166 genomes), betalactone (105 genomes), terpene (86 genomes), RiPP-like (82 genomes), siderophore (77 genomes), redox-cofactor (73 genomes), and arylpolyene (61 genomes).

Other BGCs found in Clade 6 include NRPS-like, T1PKS, hserlactone, NRPS, RRE containing, T3PKS, Lassopeptide, Ranthipeptide, resorcinol, thiopeptide, LAP, NAGGN, acyl amino acids, butyrolactone, HgIE-KS, phosphonate, PUFA, proteusin, prodigiosin, thioamitides, ladderane, CDPS, Lanthipeptide-class-iv, Lanthipeptide-

class-i, NAPAA, oligosaccharide, phenazine, PKS-like, Thioamide-NRP, transAT-PKS, and PBDE.

Clade 6 revealed that *Gynuella sunshinyii* and *Haliangium ochraceum* have a significantly higher number of biosynthetic gene clusters (BGCs) compared to other species within the clade, with 33 BGCs and 30 BGCs, respectively. These findings suggest that these species may have a greater potential for producing bioactive compounds and could be promising candidates for further investigation in natural product discovery.

The biosynthetic gene cluster of haliamide, a new PKS-NRPS hybrid metabolite, was characterized in the marine myxobacterium *Haliangium ochraceum* (Sun *et al.*, 2016).

Gynuella sunshinyii is a plant-associated marine bacterium belonging to the order Oceanospirillales that has been identified as a chemically unexplored bacterial taxon with high numbers of biosynthetic gene clusters. *In silico* chemical prediction of a non-canonical polyketide synthase cluster led to the discovery of janustatins, structurally unprecedented cytotoxins (Ueoka *et al.*, 2022).

However, some species in Clade 6, including *Desulfovibrio gracilis*, *Salinimonas chungwhensis*, *Salinimonas iocasae*, *Mycoplasma marinum*, and *Mycoplasma todarodis*, do not carry any BGCs.

The absence of biosynthetic gene clusters (BGCs) in these species may be due to several reasons. One possibility is that they have a reduced genome size and a limited metabolic capacity, which may have resulted in the loss of genes encoding secondary metabolites. Another possibility is that they rely on other mechanisms for survival and competition, such as adhesion or toxin production, and do not require the production of bioactive compounds. Finally, it is also possible that the current annotation of their genome is incomplete, and further analysis may reveal the presence of BGCs that were not initially identified. (Table 5.3)

Species	Genome availability	Genome size
<i>Desulfovibrio gracilis</i>	Contig	3.17948
<i>Mycoplasma marinum</i>	Contig	1.17115
<i>Mycoplasma todarodis</i>	Contig	1.00788
<i>Salinimonas iocasae</i>	complete genome	4.15948
<i>Salinimonas chungwhensis</i>	Scaffold	3.98544

Table 5.3 Genome Availability and Size of Halophilic Species Lacking Biosynthetic Gene Clusters (BGCs)

Also, Fig 5.5 indicates a moderate positive relationship between genome size and the number of BGCs present in halophilic bacterial genomes. This suggests that the larger the genome size of the halophilic bacteria, the greater the number of BGCs present in the genome. However, it is important to note that there are exceptions to this trend, with some genomes having a smaller size but a higher number of BGCs and vice versa. Additionally, other factors such as evolutionary history, habitat, and lifestyle of the bacteria may also contribute to the variation in BGC numbers among the genomes. Nonetheless, this finding provides insight into the potential for exploring the biosynthetic potential of halophilic bacteria with larger genomes. Further research and analysis can be conducted to investigate the specific factors influencing the relationship between genome size and BGC number in halophilic bacteria.

5.3 Future Scope

Further research is necessary to fully understand the biotechnological uses and medicinal applications of the biosynthetic gene clusters identified in halophilic bacteria. One avenue for future work could involve additional genome sequencing of the identified bacteria to gain a more complete understanding of the genes and pathways involved in the biosynthesis of potentially bioactive compounds (Scherlach & Hertweck, 2021).

Functional analysis of the identified gene clusters will also be an important aspect of future research. This could involve genetic engineering to activate or silence specific genes within the clusters, as well as biochemical assays to determine the activity of the enzymes involved in biosynthesis. Such functional analyses will be necessary to determine which biosynthetic gene clusters have the greatest potential for producing bioactive compounds and should be prioritized for further investigation (Galal *et al.*, 2023).

The identification of biosynthetic gene clusters in halophilic bacteria is just the first step in understanding their potential biotechnological and medicinal applications. Additional work is required to identify and characterize the specific compounds produced by the gene clusters and to assess their biological activity. This could involve a variety of methods, such as bioassays to determine antibacterial or antifungal activity, or pharmacological assays to assess potential uses in medicine (Pai *et al.*, 2022).

By continuing to explore the diversity and distribution of biosynthetic gene clusters in halophilic bacteria and characterizing the compounds they produce, we can better understand the unique adaptations and potential benefits of these bacteria and potentially identify new leads for biotechnological and medicinal applications.

CHAPTER 6

SUMMARY AND CONCLUSION

This project aimed to explore the diversity of biosynthetic gene clusters in halophilic bacteria through *in silico* genome mining. By analysing publicly available genomic data of halophilic bacteria using bioinformatics tools and techniques, biosynthetic gene clusters that may be involved in the production of bioactive compounds were identified. The results revealed that halophilic bacteria harbor a diverse range of biosynthetic gene clusters, with different clades and phyla having varying numbers of BGCs. The identification and characterization of biosynthetic gene clusters in halophilic bacteria is an important area of research for the discovery of novel bioactive compounds with potential biotechnological and medicinal applications.

These findings have significant implications for the field of drug discovery and biotechnology, as the compounds produced by halophilic bacteria have the potential to be used in the development of novel drugs and biotechnological solutions. Our project contributes to the growing body of knowledge about halophilic bacteria and their biosynthetic gene clusters, and provides a foundation for future research in this field.

However, it is important to note that our project is limited by the availability and quality of publicly available genomic data. Further research involving more extensive genome sequencing and functional analysis of the identified biosynthetic gene clusters is necessary to fully understand their potential biotechnological uses and medicinal applications.

In summary, this project highlights the potential of halophilic bacteria and their biosynthetic gene clusters as a source of novel bioactive compounds, and provides a starting point for future research in this field.

REFERENCES

1. Abed, R.M.M., S. Dobretsov, and K. Sudesh. "Applications of Cyanobacteria in Biotechnology." *Journal of Applied Microbiology* 106, no. 1 (January 2009): 1–12. <https://doi.org/10.1111/j.1365-2672.2008.03918.x>.
2. Alam, Khorshed, Md Mahmudul Islam, Kai Gong, Muhammad Nazeer Abbasi, Ruijuan Li, Youming Zhang, and Aiyang Li. "In Silico Genome Mining of Potential Novel Biosynthetic Gene Clusters for Drug Discovery from Burkholderia Bacteria." *Computers in Biology and Medicine* 140 (January 2022): 105046. <https://doi.org/10.1016/j.compbiomed.2021.105046>.
3. Albarano, Luisa, Roberta Esposito, Nadia Ruocco, and Maria Costantini. "Genome Mining as New Challenge in Natural Products Discovery." PubMed Central (PMC), April 9, 2020. <https://doi.org/10.3390/md18040199>.
4. Aljohny, Bassam Oudh. "Halophilic Bacterium – A Review of New Studies – Biosciences Biotechnology Research Asia." Halophilic Bacterium – A Review of New Studies – Biosciences Biotechnology Research Asia, n.d. <https://www.biotech-asia.org/vol12no3/halophilic-bacterium-a-review-of-new-studies/>.
5. Anand, Swadha, M. V. R. Prasad, Gitanjali Yadav, Narendra Kumar, Jyoti Shehara, Md. Zeeshan Ansari, and Debasisa Mohanty. "SBSPKS: Structure Based Sequence Analysis of Polyketide Synthases." PubMed Central (PMC), May 5, 2010. <https://doi.org/10.1093/nar/gkq340>.
6. Atanasov, Atanas G., Sergey B. Zotchev, Verena M. Dirsch, and Claudiu T. Supuran. "Natural Products in Drug Discovery: Advances and Opportunities - Nature Reviews Drug Discovery." *Nature*, January 28, 2021. <https://doi.org/10.1038/s41573-020-00114-z>.
7. Barka, Essaid Ait, Parul Vatsa, Lisa Sanchez, Nathalie Gaveau-Vaillant, Cedric Jacquard, Hans-Peter Klenk, Christophe Clément, Yder Ouhdouch, and Gilles P. van Wezel. "Taxonomy, Physiology, and Natural Products of Actinobacteria." *Microbiology and Molecular Biology Reviews* 80, no. 1 (March 2016): 1–43. <https://doi.org/10.1128/mmbr.00019-15>.
8. Bauman, Katherine D., Keelie S. Butler, Bradley S. Moore, and Jonathan R. Chekan. "Genome Mining Methods to Discover Bioactive Natural Products."

- PubMed Central (PMC), November 4, 2021. <https://doi.org/10.1039/d1np00032b>.
9. Bennur, Tahsin, Ameeta Ravi Kumar, Smita Zinjarde, and Vaishali Javdekar. "Nocardiosis Species: Incidence, Ecological Roles and Adaptations." *Microbiological Research* 174 (May 2015): 33–47. <https://doi.org/10.1016/j.micres.2015.03.010>.
 10. Blin, Kai, Hyun Uk Kim, Marnix H Medema, and Tilmann Weber. "Recent Development of antiSMASH and Other Computational Approaches to Mine Secondary Metabolite Biosynthetic Gene Clusters." PubMed Central (PMC), November 3, 2017. <https://doi.org/10.1093/bib/bbx146>.
 11. Blin, Kai, Marnix H. Medema, Daniyal Kazempour, Michael A. Fischbach, Rainer Breitling, Eriko Takano, and Tilmann Weber. "antiSMASH 2.0—a Versatile Platform for Genome Mining of Secondary Metabolite Producers." OUP Academic, July 1, 2013. <https://doi.org/10.1093/nar/gkt449>.
 12. Blin, Kai, Simon Shaw, Katharina Steinke, Rasmus Villebro, Nadine Ziemert, Sang Yup Lee, Marnix H Medema, and Tilmann Weber. "antiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline." OUP Academic, July 2, 2019. <https://doi.org/10.1093/nar/gkz310>.
 13. Blin, Kai, Victòria Pascal Andreu, Emmanuel L C de los Santos, Francesco Del Carratore, Sang Yup Lee, Marnix H Medema, and Tilmann Weber. "The antiSMASH Database Version 2: A Comprehensive Resource on Secondary Metabolite Biosynthetic Gene Clusters." PubMed Central (PMC), November 5, 2018. <https://doi.org/10.1093/nar/gky1060>.
 14. Carver, Tim, Simon R. Harris, Matthew Berriman, Julian Parkhill, and Jacqueline A. McQuillan. "Artemis: An Integrated Platform for Visualization and Analysis of High-Throughput Sequence-Based Experimental Data." PubMed Central (PMC), December 22, 2011. <https://doi.org/10.1093/bioinformatics/btr703>.
 15. Chen, Ray, Hon Lun Wong, Gareth S. Kindler, Fraser Iain MacLeod, Nicole Benaud, Belinda C. Ferrari, and Brendan P. Burns. "Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats." *Frontiers*, July 24, 2020. <https://doi.org/10.3389/fmicb.2020.01950>.
 16. Chen, Yan-Huey, Chia-Wei Lu, Yuan-Tay Shyu, and Shih-Shun Lin. "Revealing the Saline Adaptation Strategies of the Halophilic Bacterium

- Halomonas Beimenensis through High-Throughput Omics and Transposon Mutagenesis Approaches - Scientific Reports.” Nature, October 12, 2017. <https://doi.org/10.1038/s41598-017-13450-9>.
17. Cimermancic, Peter, Marnix H. Medema, Jan Claesen, Kenji Kurita, Laura C. Wieland Brown, Konstantinos Mavrommatis, Amrita Pati, et al. “Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters.” PubMed Central (PMC), n.d. <https://doi.org/10.1016/j.cell.2014.06.034>.
 18. Conde-Martínez, Natalia, Alejandro Acosta-González, Luis E. Díaz, and Edison Tello. “Use of a Mixed Culture Strategy to Isolate Halophilic Bacteria with Antibacterial and Cytotoxic Activity from the Manaure Solar Saltern in Colombia.” *BMC Microbiology* 17, no. 1 (December 2017). <https://doi.org/10.1186/s12866-017-1136-x>.
 19. Corral, Paulina, Mohammad A. Amoozegar, and Antonio Ventosa. “Halophiles and Their Biomolecules: Recent Advances and Future Applications in Biomedicine.” PubMed Central (PMC), December 30, 2019. <https://doi.org/10.3390/md18010033>.
 20. Cortes, Jesus, Stephen F. Haydock, Gareth A. Roberts, Debra J. Bevitt, and Peter F. Leadlay. “An Unusually Large Multifunctional Polypeptide in the Erythromycin-Producing Polyketide Synthase of *Saccharopolyspora erythraea* - Nature.” Nature, November 8, 1990. <https://doi.org/10.1038/348176a0>.
 21. DasSarma, Satyajit L, Melinda D Capes, Priya DasSarma, and Shiladitya DasSarma. “HaloWeb: The Haloarchaeal Genomes Database.” PubMed Central (PMC), December 30, 2010. <https://doi.org/10.1186/1746-1448-6-12>.
 22. Demain, Arnold L. “Importance of Microbial Natural Products and the Need to Revitalize Their Discovery.” *Journal of Industrial Microbiology and Biotechnology* 41, no. 2 (February 1, 2014): 185–201. <https://doi.org/10.1007/s10295-013-1325-z>.
 23. Dias, Daniel A., Sylvia Urban, and Ute Roessner. “A Historical Overview of Natural Products in Drug Discovery.” PubMed Central (PMC), April 16, 2012. <https://doi.org/10.3390/metabo2020303>.

24. Duban, Matthieu, Stéphane Cociancich, and Valérie Leclère. "Nonribosomal Peptide Synthesis Definitely Working Out of the Rules." PubMed Central (PMC), March 7, 2022. <https://doi.org/10.3390/microorganisms10030577>.
25. Dutta, Bhramar, and Rajib Bandopadhyay. "Biotechnological Potentials of Halophilic Microorganisms and Their Impact on Mankind - Beni-Suef University Journal of Basic and Applied Sciences." SpringerOpen, May 31, 2022. <https://doi.org/10.1186/s43088-022-00252-w>.
26. Edbeib, Mohamed Faraj, Roswanira Abdul Wahab, and Fahrul Huyop. "Halophiles: Biology, Adaptation, and Their Role in Decontamination of Hypersaline Environments - World Journal of Microbiology and Biotechnology." SpringerLink, June 25, 2016. <https://doi.org/10.1007/s11274-016-2081-9>.
27. Forli, Stefano. "Epothilones: From Discovery to Clinical Trials." PubMed Central (PMC), n.d. <https://doi.org/10.2174/1568026614666141130095855>.
28. Galal, Aya, Salma Abou Elhassan, Ahmed H. Saleh, Ahmed I. Ahmed, Mennatallah M. Abdelrahman, Michael M. Kamal, Rana S. Khalel, and Laila Ziko. "A Survey of the Biosynthetic Potential and Specialized Metabolites of Archaea and Understudied Bacteria." *Current Research in Biotechnology* 5 (2023): 100117. <https://doi.org/10.1016/j.crbiot.2022.11.004>.
29. Gunde-Cimerman, Nina, Ana Plemenitaš, and Aharon Oren. "Strategies of Adaptation of Microorganisms of the Three Domains of Life to High Salt Concentrations." *FEMS Microbiology Reviews* 42, no. 3 (February 26, 2018): 353–75. <https://doi.org/10.1093/femsre/fuy009>.
30. Hammami, Riadh, Abdelmajid Zouhir, Christophe Le Lay, Jeannette Ben Hamida, and Ismail Fliss. "BACTIBASE Second Release: A Database and Tool Platform for Bacteriocin Characterization." PubMed Central (PMC), January 27, 2010. <https://doi.org/10.1186/1471-2180-10-22>.
31. Hannigan, Geoffrey D, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, et al. "Deep Learning Genome-Mining Strategy for Biosynthetic Gene Cluster Prediction." OUP Academic, October 10, 2019. <https://doi.org/10.1093/nar/gkz654>.
32. HARVEY, A. "Natural Products in Drug Discovery." *Drug Discovery Today* 13, no. 19–20 (October 2008): 894–901. <https://doi.org/10.1016/j.drudis.2008.07.004>.

33. Huang, Kai, Bo Zhang, Zhen-Yang Shen, Xue Cai, Zhi-Qiang Liu, and Yu-Guo Zheng. "Enhanced Amphotericin B Production by Genetically Engineered *Streptomyces Nodosus*." *Microbiological Research* 242 (January 2021): 126623. <https://doi.org/10.1016/j.micres.2020.126623>.
34. Ichikawa, Natsuko, Machi Sasagawa, Mika Yamamoto, Hisayuki Komaki, Yumi Yoshida, Shuji Yamazaki, and Nobuyuki Fujita. "DoBISCUIT: A Database of Secondary Metabolite Biosynthetic Gene Clusters." PubMed Central (PMC), November 25, 2012. <https://doi.org/10.1093/nar/gks1177>.
35. Jose, Polpass Arul, Anjisha Maharshi, and Bhavanath Jha. "Actinobacteria in Natural Products Research: Progress and Prospects." *Microbiological Research* 246 (May 2021): 126708. <https://doi.org/10.1016/j.micres.2021.126708>.
36. Kautsar, Satria A, Justin J J van der Hooft, Dick de Ridder, and Marnix H Medema. "BiG-SLiCE: A Highly Scalable Tool Maps the Diversity of 1.2 Million Biosynthetic Gene Clusters." OUP Academic, January 1, 2021. <https://doi.org/10.1093/gigascience/giaa154>.
37. Kautsar, Satria A, Kai Blin, Simon Shaw, Tilmann Weber, and Marnix H Medema. "BiG-FAM: The Biosynthetic Gene Cluster Families Database." PubMed Central (PMC), October 3, 2020. <https://doi.org/10.1093/nar/gkaa812>.
38. LATHAN, BERND, and DANIEL D. VON HOFF. "Cytotoxic Activity of Echinomycin in a Human Tumor Cloning System." *Cancer Drug Delivery* 1, no. 3 (January 1984): 191–98. <https://doi.org/10.1089/cdd.1984.1.191>.
39. Lebedeva, Jolanta, Gabriele Juknevičute, Rimvydė Čepaitė, Vida Vickackaite, Raminta Pranckutė, and Nomeda Kuisiene. "Genome Mining and Characterization of Biosynthetic Gene Clusters in Two Cave Strains of *Paenibacillus* Sp." *Frontiers*, December 11, 2020. <https://doi.org/10.3389/fmicb.2020.612483>.
40. Li, Lei, Logan W. MacIntyre, and Sean F. Brady. "Refactoring Biosynthetic Gene Clusters for Heterologous Production of Microbial Natural Products." PubMed Central (PMC), January 18, 2021. <https://doi.org/10.1016/j.copbio.2020.12.011>.
41. Li, Michael HT, Peter MU Ung, James Zajkowski, Sylvie Garneau-Tsodikova, and David H Sherman. "Automated Genome Mining for Natural Products -

- BMC Bioinformatics.” BioMed Central, June 16, 2009. <https://doi.org/10.1186/1471-2105-10-185>.
42. Liang, Haoyu, Zhiman Song, Zheng Zhong, Dengwei Zhang, Wei Yang, Le Zhou, Ethan Older, et al. “Mining of the Biosynthetic Potential of Archaea Reveals Diverse Anti-Archaeal Lanthipeptides,” March 23, 2022. <https://doi.org/10.21203/rs.3.rs-1476892/v1>.
 43. Loukas, Alexios, Ilias Kappas, and Theodore J. Abatzopoulos. “HaloDom: A New Database of Halophiles across All Life Domains.” PubMed Central (PMC), January 15, 2018. <https://doi.org/10.1186/s40709-017-0072-0>.
 44. Matter, Andrea M., Sara B. Hoot, Patrick D. Anderson, Susana S. Neves, and Yi-Qiang Cheng. “Valinomycin Biosynthetic Gene Cluster in Streptomyces: Conservation, Ecology and Evolution.” PubMed Central (PMC), September 29, 2009. <https://doi.org/10.1371/journal.pone.0007194>.
 45. Medema, Marnix H, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene de Bruijn, et al. “Minimum Information about a Biosynthetic Gene Cluster - Nature Chemical Biology.” Nature, August 18, 2015. <https://doi.org/10.1038/nchembio.1890>.
 46. Miao, Vivian, Marie-Françoise Coëffet-LeGal, Paul Brian, Renee Brost, Julia Penn, Andrew Whiting, Steven Martin, et al. “Daptomycin Biosynthesis in Streptomyces Roseosporus: Cloning and Analysis of the Gene Cluster and Revision of Peptide Stereochemistry.” microbiologyresearch.org, May 1, 2005. <https://doi.org/10.1099/mic.0.27757-0>.
 47. Mohammadipanah, Fatemeh, Javad Hamedi, and Mona Dehghani. “Halophilic Bacteria: Potentials and Applications in Biotechnology.” *Sustainable Development and Biodiversity*, 2015, 277–321. https://doi.org/10.1007/978-3-319-14595-2_11.
 48. Mungan, Mehmet Direnç, Mohammad Alanjary, Kai Blin, Tilmann Weber, Marnix H Medema, and Nadine Ziemert. “ARTS 2.0: Feature Updates and Expansion of the Antibiotic Resistant Target Seeker for Comparative Genome Mining.” PubMed Central (PMC), May 19, 2020. <https://doi.org/10.1093/nar/gkaa374>.
 49. Navarro-Muñoz, Jorge C., Nelly Selem-Mojica, Michael W. Mullowney, Satria A. Kautsar, James H. Tryon, Elizabeth I. Parkinson, Emmanuel L.C. De Los Santos, et al. “A Computational Framework to Explore Large-Scale

- Biosynthetic Diversity.” PubMed Central (PMC), November 25, 2019. <https://doi.org/10.1038/s41589-019-0400-9>.
50. Ordóñez-Robles, María, Fernando Santos-Beneit, and Juan F. Martín. “Unraveling Nutritional Regulation of Tacrolimus Biosynthesis in *Streptomyces Tsukubaensis* through Omic Approaches.” MDPI, May 1, 2018. <https://doi.org/10.3390/antibiotics7020039>.
 51. Pai, Sanidhya, Akshatha Hebbar, and Subbalaxmi Selvaraj. “A Critical Look at Challenges and Future Scopes of Bioactive Compounds and Their Incorporations in the Food, Energy, and Pharmaceutical Sector - Environmental Science and Pollution Research.” SpringerLink, March 2, 2022. <https://doi.org/10.1007/s11356-022-19423-4>.
 52. Palaniappan, Krishnaveni, I-Min A Chen, Ken Chu, Anna Ratner, Rekha Seshadri, Nikos C Kyrpides, Natalia N Ivanova, and Nigel J Mouncey. “IMG-ABC v.5.0: An Update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase.” PubMed Central (PMC), October 29, 2019. <https://doi.org/10.1093/nar/gkz932>.
 53. Park, Sung Ryeol, Young Ji Yoo, Yeon-Hee Ban, and Yeo Joon Yoon. “Biosynthesis of Rapamycin and Its Regulation: Past Achievements and Recent Progress - The Journal of Antibiotics.” Nature, June 30, 2010. <https://doi.org/10.1038/ja.2010.71>.
 54. PETKOVIC, HRVOJE, RACHEL E. LILL, ROSE M. SHERIDAN, BARRIE WILKINSON, ELLEN L. McCORMICK, HAMISH A. I. McARTHUR, JAMES STAUNTON, PETER F. LEADLAY, and STEVEN G. KENDREW. “A Novel Erythromycin, 6-Desmethyl Erythromycin D, Made by Substituting an Acyltransferase Domain of the Erythromycin Polyketide Synthase.” A Novel Erythromycin, 6-Desmethyl Erythromycin D, Made by Substituting an Acyltransferase Domain of the Erythromycin Polyketide Synthase, January 27, 2009. <https://doi.org/10.7164/antibiotics.56.543>.
 55. Ramos-Cormenzana, Alberto. “Halophilic Organisms and the Environment.” *General and Applied Aspects of Halophilic Microorganisms*, 1991, 15–24. https://doi.org/10.1007/978-1-4615-3730-4_3.
 56. Risdian, Chandra, Tjandrawati Mozef, and Joachim Wink. “Biosynthesis of Polyketides in *Streptomyces*.” PubMed Central (PMC), May 6, 2019. <https://doi.org/10.3390/microorganisms7050124>.

57. Russell, Alicia H., and Andrew W. Truman. "Genome Mining Strategies for Ribosomally Synthesised and Post-Translationally Modified Peptides." PubMed Central (PMC), June 25, 2020. <https://doi.org/10.1016/j.csbj.2020.06.032>.
58. Sasidharan, S, Y Chen, D Saravanan, K M Sundram, and L Yoga Latha. "Extraction, Isolation and Characterization of Bioactive Compounds from Plants' Extracts." PubMed Central (PMC), October 2, 2010. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218439/>.
59. Scherlach, Kirstin, and Christian Hertweck. "Mining and Unearthing Hidden Biosynthetic Potential - Nature Communications." Nature, June 23, 2021. <https://doi.org/10.1038/s41467-021-24133-5>.
60. Sélem-Mojica, Nelly, César Aguilar, Karina Gutiérrez-García, Christian E. Martínez-Guerrero, and Fancisco Barona-Gómez. "EvoMining Reveals the Origin and Fate of Natural Product Biosynthetic Enzymes." PubMed Central (PMC), April 4, 2019. <https://doi.org/10.1099/mgen.0.000260>.
61. Sharma, Naveen, Mohammad Samir Farooqi, Krishna Kumar Chaturvedi, Shashi Bhushan Lal, Monendra Grover, Anil Rai, and Pankaj Pandey. "The Halophile Protein Database." PubMed Central (PMC), December 1, 2014. <https://doi.org/10.1093/database/bau114>.
62. Shetty, Prakasham Reddy, Sudheer Kumar Buddana, Vinay Bharadwaj Tatipamula, Yaswanth Varanasi Venkata Naga, and Jamal Ahmad. "Production of Polypeptide Antibiotic from Streptomyces Parvulus and Its Antibacterial Activity." PubMed Central (PMC), April 8, 2014. <https://doi.org/10.1590/S1517-83822014005000022>.
63. Skinnider, Michael A., Chad W. Johnston, Mathusan Gunabalasingam, Nishanth J. Merwin, Agata M. Kieliszek, Robyn J. MacLellan, Haoxin Li, et al. "Comprehensive Prediction of Secondary Metabolite Structure and Biological Activity from Microbial Genome Sequences - Nature Communications." Nature, November 27, 2020. <https://doi.org/10.1038/s41467-020-19986-1>.
64. Sooklal, Selisha Ann, Charles De Koning, Dean Brady, and Karl Rumbold. "Identification and Characterisation of a Fluorinase from Actinopolyspora Mzabensis." *Protein Expression and Purification* 166 (February 2020): 105508. <https://doi.org/10.1016/j.pep.2019.105508>.

65. Starcevic, Antonio, Jurica Zucko, Jurica Simunkovic, Paul F. Long, John Cullum, and Daslav Hranueli. "ClustScan : An Integrated Program Package for the Semi-Automatic Annotation of Modular Biosynthetic Gene Clusters and in Silico Prediction of Novel Chemical Structures." *Nucleic Acids Research* 36, no. 21 (October 31, 2008): 6882–92. <https://doi.org/10.1093/nar/gkn685>.
66. Sun, Yuwei, Tomohiko Tomura, Junichi Sato, Takashi Iizuka, Ryosuke Fudou, and Makoto Ojika. "Isolation and Biosynthetic Analysis of Haliamide, a New PKS-NRPS Hybrid Metabolite from the Marine Myxobacterium *Haliangium ochraceum*." *Molecules* 21, no. 1 (January 6, 2016): 59. <https://doi.org/10.3390/molecules21010059>.
67. Terlouw, Barbara R, Kai Blin, Jorge C Navarro-Muñoz, Nicole E Avalon, Marc G Chevrette, Susan Egbert, Sanghoon Lee, et al. "MIBiG 3.0: A Community-Driven Effort to Annotate Experimentally Validated Biosynthetic Gene Clusters." PubMed Central (PMC), November 18, 2022. <https://doi.org/10.1093/nar/gkac1049>.
68. Ueoka, Reiko, Philipp Sonderrmann, Stefan Leopold-Messer, Yizhou Liu, Rei Suo, Agneya Bhushan, Lida Vadakumchery, et al. "Genome-Based Discovery and Total Synthesis of Janustatins, Potent Cytotoxins from a Plant-Associated Bacterium." PubMed Central (PMC), September 5, 2022. <https://doi.org/10.1038/s41557-022-01020-0>.
69. Ukani, Hetal, Megha K Purohit, John J Georrgie, Sneha Paul, and Satya P Singh. "HaloBase: Development of Database System for Halophilic Bacteria and Archaea with Respect to Proteomics, Genomics & Other Molecular Traits." Zenodo, November 1, 2011. <https://doi.org/10.5281/zenodo.4398512>.
70. van Heel, Auke J, Anne de Jong, Chunxu Song, Jakob H Viel, Jan Kok, and Oscar P Kuipers. "BAGEL4: A User-Friendly Web Server to Thoroughly Mine RiPPs and Bacteriocins." OUP Academic, July 2, 2018. <https://doi.org/10.1093/nar/gky383>.
71. Ventosa, Antonio, Joaquín J. Nieto, and Aharon Oren. "Biology of Moderately Halophilic Aerobic Bacteria." PubMed Central (PMC), n.d. <https://doi.org/10.1128/mmbr.62.2.504-544.1998>.

72. Vijayakumar, Ramasamy, and Suresh S. S. Raja. *Secondary Metabolites: Sources and Applications*, 2018.
73. Viju, Nadarajan, Stanislaus Mary Josephine Punitha, and Sathianeson Satheesh. "An Analysis of Biosynthesis Gene Clusters and Bioactivity of Marine Bacterial Symbionts." *Current Microbiology* 78, no. 7 (May 26, 2021): 2522–33. <https://doi.org/10.1007/s00284-021-02535-4>.
74. Vishwakarma, Rashi, and Ashwani Kumar Rai. "Separation of Bioactive Metabolites from *Aphanothece Halophytica* Through HPLC and Characterization of the Analytes Through ESI-MS and NMR." *The Natural Products Journal* 3, no. 2 (May 1, 2013): 151–57. <https://doi.org/10.2174/2210315511303020010>.
75. Vishwakarma, Rashi. "Inhibitory Microbial Activity and GC-MS Based Metabolite Profile of the Halophilic Cyanobacterium *Aphanothece Halophytica*." *Algological Studies* 141 (April 1, 2013): 73–91. <https://doi.org/10.1127/1864-1318/2013/0101>.
76. Waditee-Sirisattha, Rungaroon, Hakuto Kageyama, and Teruhiro Takabe. "Halophilic Microorganism Resources and Their Applications in Industrial and Environmental Biotechnology." *AIMS Microbiology*, n.d. <https://doi.org/10.3934/microbiol.2016.1.42>.
77. Wang, Jia, Ruihua Zhang, Xin Chen, Xinxiao Sun, Yajun Yan, Xiaolin Shen, and Qipeng Yuan. "Biosynthesis of Aromatic Polyketides in Microorganisms Using Type II Polyketide Synthases - Microbial Cell Factories." *BioMed Central*, May 24, 2020. <https://doi.org/10.1186/s12934-020-01367-4>.
78. Xu, Dongbo, Linna Han, Chunhui Li, Qi Cao, Duolong Zhu, Nolan H. Barrett, Dedra Harmody, et al. "Bioprospecting Deep-Sea Actinobacteria for Novel Anti-Infective Natural Products." *PubMed Central (PMC)*, April 30, 2018. <https://doi.org/10.3389/fmicb.2018.00787>.
79. Yang, Xiuqing, Peng Feng, Ying Yin, Kathryn Bushley, Joseph W. Spatafora, and Chengshu Wang. "Cyclosporine Biosynthesis in *Tolypocladium Inflatum* Benefits Fungal Adaptation to the Environment." Edited by B. Gillian Turgeon. *mBio* 9, no. 5 (November 7, 2018). <https://doi.org/10.1128/mbio.01211-18>.
80. Yang, Xiuqing, Peng Feng, Ying Yin, Kathryn Bushley, Joseph W. Spatafora, and Chengshu Wang. "Cyclosporine Biosynthesis in *Tolypocladium Inflatum*

- Benefits Fungal Adaptation to the Environment.” Edited by B. Gillian Turgeon. *mBio* 9, no. 5 (November 7, 2018). <https://doi.org/10.1128/mbio.01211-18>.
81. Yang, Xu, and Ahmed E. Yousef. “Antimicrobial Peptides Produced by *Brevibacillus* Spp.: Structure, Classification and Bioactivity: A Mini Review - World Journal of Microbiology and Biotechnology.” SpringerLink, March 29, 2018. <https://doi.org/10.1007/s11274-018-2437-4>.
 82. Yin, Ying, Menghao Cai, Xiangshan Zhou, Zhiyong Li, and Yuanxing Zhang. “Polyketides in *Aspergillus Terreus*: Biosynthesis Pathway Discovery and Application - Applied Microbiology and Biotechnology.” SpringerLink, July 26, 2016. <https://doi.org/10.1007/s00253-016-7733-z>.
 83. Zhang, Qing-Wen, Li-Gen Lin, and Wen-Cai Ye. “Techniques for Extraction and Isolation of Natural Products: A Comprehensive Review - Chinese Medicine.” SpringerLink, April 17, 2018. <https://doi.org/10.1186/s13020-018-0177-x>.
 84. Zhu, Jiang, Shiyi Wang, Cheng Wang, Zhi Wang, Gan Luo, Junhui Li, Yangyang Zhan, Dongbo Cai, and Shouwen Chen. “Microbial Synthesis of Bacitracin: Recent Progress, Challenges, and Prospects.” *Synthetic and Systems Biotechnology* 8, no. 2 (June 2023): 314–22. <https://doi.org/10.1016/j.synbio.2023.03.009>.
 85. Ziemert, Nadine, Sheila Podell, Kevin Penn, Jonathan H. Badger, Eric Allen, and Paul R. Jensen. “The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity.” The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity | PLOS ONE, March 29, 2012. <https://doi.org/10.1371/journal.pone.0034064>.