THE AMERICAN UNIVERSITY IN CAIRO
الجامعة الأمريكية بالقاهرة

School of Sciences and Engineering

Mining selected metagenomes/metatranscriptomes for biosynthetic gene clusters and antimicrobial resistance genes

A Thesis Submitted to

The Biotechnology Master's Program

In partial fulfilment of the requirements for

The degree of Master of Science

By: Ahmed Yamany Abdel Aziz

Under the supervision of:

Prof. Hassan Azzazy, Prof. Ahmed Moustafa and Assist. Prof. Laila Ziko

American University in Cairo

Fall / 2020

The American University in Cairo

Mining selected metagenomes/metatranscriptomes for biosynthetic gene clusters and antimicrobial resistance gene

A Thesis Submitted by

Ahmed Yamany Abdel Aziz

To the Biotechnology Graduate Program

Fall / 2020

In partial fulfillment of the requirements for

The degree of Master of Science

Has been approved by

Thesis Committee Supervisor/Chair

_____

Affiliation _____

Thesis Committee Reader/Examiner

_____

Affiliation _____

Thesis Committee Reader/Examiner

_____

Affiliation _____

Thesis Committee Reader/External Examiner

_____

Affiliation_____

_____ _____ _____ _____

Dept. Chair/Director Date Dean Date

Dedication and Acknowledgements

I would like to express my immense appreciation and gratitude to my advisors Professor Hassan Azzazy, Professor Ahmed Moustafa and Assistant Professor Laila Ziko for giving me the opportunity to join their research team. It would not have been possible to accomplish my thesis research without Prof. Azzazy's expert advice, continuous support, critical discussion, and endless patience. A very special thanks for my advisor, mentor, facilitator, big brother, Prof. Ahmed Moustafa, I am extremely blessed having him as my research advisor. I cannot thank him enough for providing a healthy working environment. I would also like to thank Dr. Laila Ziko, for her thorough insights, invaluable suggestions in daily basis, and being a constant source of encouragement and motivation, without her support and guidance I will never reach this point.

Special thanks goes to the examination committee, Dr. Khaled Abou Aisha, Dr. Ahmed Abdel-Latif and Dr. Arthur Bos for giving me their invaluable advices.

Grateful thanks are extended to my old friends and lab mates, Ahmed Elbaz, Abu Baker, Amgad, Ahmed Housseiny and Moustafa for their valuable help, encouragement, and providing friendly and cooperative work environment. A special thanks to my dearest friend and Brother Dr. Ibrahim Farag, for his generous support and care during my MSc. Dr. Farag was the one who advised me to join the program and he was behind any single success during this journey.

I would like to send my sincere gratitude to my family: my parents and my siblings for supporting me throughout my MSc Journey.

Last but not least, I would like to express my deepest appreciation to my beloved wife for her unconditional love, endless support, and encouragement. Also, I would like to offer the fruit of this work to my four precious sons, Ammar, Walid, Anas and Mohamed, for all the patience and love they showed.

To you all, I dedicate this thesis, which I wish to help decrease the burden caused by antibiotic resistance worldwide one day.

Abstract

Antimicrobial resistance is one of the serious global challenges in the current century. The fact that resistance genes transfer between bacteria, coupled with the fact that the world is connected through complex dynamics. Studying microbial behavior and understanding the different factors coffering microbial resistance to a broad spectrum of the available drug classes, parallel with a comprehensive analysis of the natural microbial products as the primary source of the novel antibiotics, might shed some light on solutions for this problem.

Microbial environments harbor a wide range of secondary metabolites (SM) with different functional groups. SMs are not directly involved in vital microbial processes such as reproduction, growth, and development. However, these organic compounds, which exist in many different chemical structures, carry out a broad range of functions. Some bioactive SMs are widely used in drug development of various therapeutic classes such as antibacterial, anticancer, immunosuppressant, diabetic, and cholesterol-lowering agents. These bioactive compounds' metabolic pathways are encoded by co-localized genes collectively called Biosynthetic Gene Clusters (BGCs). The majority of the discovered bioactive natural products are from microbial strains that are cultivatable. However, the advancement in sequencing techniques, bioinformatics, and metagenomics opened unlimited opportunities to reach and study the uncultivatable microbial communities, which represent the more significant fraction of the underexplored microbial ecology.

In this study, selected samples of seven selected metatranscriptomic/metagenomic datasets were subjected to assembly, taxonomic assignment to the reads, and assembled contigs. The aim of this study is two-fold. Firstly, the assembled contigs were then investigated by two primary distinct computational methods, namely antibiotics and Secondary Metabolite Analysis Shell (antiSMASH) and deep-learning (deepBGC) methods. A comparative study was performed to determine the biosynthetic gene clusters (BGCs) present in each of the included samples and compare their taxonomic differences. Secondly, the assembled contigs were also analyzed to determine the antimicrobial resistance (AMR) genes present in each sample by using the Resistance Gene Identifier (RGI) algorithm, which is a part of the Comprehensive Antibiotic Resistance Database (CARD). A total of 65 samples from the seven selected

metagenomic and metatranscriptomic datasets were investigated by antiSMASH, deepBGC pipelines, and CARD in the present study. The different classes of detected BGCs and their corresponding microbial taxa and the antimicrobial resistance gene families and their corresponding resistance mechanisms against specific drug classes were reported.

In the current study, we reported that the datasets with a large extent of variability (i.e. sex, age and illness state) due to the nature of their environments, such as host microbiome samples of patients in two ecosystems (COVID-19 & Atopic Dermatitis), gave the most variable number of BGC classes detected by antiSMASH, where 19 different classes detected in skin microbiome of AD patients and 16 different classes detected in gut microbiome of COVID-19 patients. On the other hand and due to the selection pressure on the microbial ecosystems by the wide use of antibiotics, gut microbiome of COVID-19 patients' and water sewage samples had more than 70% of the detected AMR gene families where gut microbiome of COVID-19 patients' sample alone reported to had more than 50% of AMR genes detected by CARD.

In conclusion, ecological characteristics and microbial diversity in terms of composition and relative abundance dramatically affect the dynamics of secondary metabolites' production and transferring antimicrobial resistance genes between bacteria. Microbial strains with higher biosynthetic and antimicrobial resistance potentials were enriched in environments with a rich microbial diversity such as host microbiome (i.e., COVID-19 patients), with patterns of abundance of biosynthetic gene clusters and AMR genes fluctuating by taxonomy.

Table of Contents

Dedication and Acknowledgments iii

Abstract iv

Table of Contents vi

Glossary and Abbreviations viii

List of Tables ix

List of Figures x

Chapter 1: Literature Review & Study Objectives 1

Introduction 1

Global health challenges associated with antibiotic and chemotherapeutic resistances 1

Natural Products and their Pharmaceutical Importance 1

Natural Products are encoded by Biosynthetic Gene Clusters (BGCs) 2

Tools for BGC mining and NP Discovery 3

NP mining: past, present and future 3

Omics Approaches in NP Discovery 3

In silico Tools for Biosynthetic Gene Clusters Identification 4

Genomics and High-throughput Sequencing Technologies 5

The importance of searching for BGCs in metagenomes 6

AntiSMASH platform to detect BGCs 7

Deep learning method to detect BGCs 8

Study Objectives 9

Chapter 2: Materials & Methods 10

Samples and Assembly 10

Taxonomic analysis, annotation and bioinformatic visualization 11

Using deepBGC tool for BGC mining 11

Using antiSMASH tool for BGC mining 12

Using CARD's RGI algorithm for AMR genes detection 12

Statistical Analyses 12
Chapter 3: Results 13
Assembly pipeline 13
Taxonomical Assignment 17
BGC profile of samples in the dataset as detected by antiSMASH 19
BGC profile of samples in the dataset as detected by DeepBGC 23
AMR genes profile of samples in the dataset as detected by CARD's RGI 28
Chapter 4: Discussion 37
Different environments have different microbial taxa profile 37
The Biosynthetic Potential of the selected microbial metagenomes 42
Comparison of BGCs as detected by DeepBGC and antiSMASH 47
AMR genes profile of samples in the dataset as detected by CARD's RGI 48
The effect of antibiotic use in AMR genes transfer between microbial communities. 50
AMR genes in gut microbiome of COVID-19 patients' samples 51
AMR genes in water sewage 52
Antibiotic efflux resistance mechanism in gut microbiome of COVID-19 patients' and water sewage samples 53
Chapter 5: Conclusions and Future Perspectives 54
Conclusions 54
Future Perspectives 57
References 59

Glossary and Abbreviations

AMR	Antimicrobial Resistance
antiSMASH	antibiotics & Secondary Metabolite Analysis Shell
BGC	Biosynthetic Gene Cluster
BiLSTM	Bidirectional Long Short-Term Memory
BLAST	Basic Local Alignment Search Tool
CARD	Comprehensive Antibiotic Resistance Database
CCS	Circular Consensus Sequencing
Contig	A set of overlapping DNA segments that together represent a consensus region of DNA (From contiguous).
hglE-KS	heterocyst glycolipid synthase-like PKS
HGT	Horizontal Gene Transfer
HMMs	Hidden Markov Models
hserlactone	Homoserine lactone cluster
LAP	Linear azol(in)e-containing peptides
MFS	Major Facilitator Superfamily
MGE	Mobile Genetic Element
MIBiG	Minimum Information about a Biosynthetic Gene Cluster
NAGGN	N-γ-acetylglutaminyl glutamine 1-amide
NGS	Next-generation Sequencing
NLP	Natural Language Processing
NP	Natural Products
NRPS	Non--ribosomal peptide synthases
NRPS-like	NRPS-like fragment
Pfam	Protein family
PKS	Polyketide synthases
PKS-like	Other types of PKS cluster
RGI	Resistance Gene Identifier
RND	Resistance-nodulation-cell division
RNNs	Recurrent Neural Networks
SM	Secondary Metabolites
SV	Structural Variation
T1PKS	Type I PKS (Polyketide synthase)
T2PKS	Type II PKS
WHO	World Health Organisation

List of Tables

Table 1. List of the widely used tools used for prediction of BGCs modified from Ren H et al., 2020. 5

Table 2. Strengths and weaknesses of some available long-read platforms, data was modified from van Dijk EL et al., 2018. 6

Table 3. The selected metagenomic and metatranscriptomic projects included in the dataset... 14

Table 4. Assembly metrics denoted for each sample from the selected metagenomic and metatranscriptomic projects included in the dataset.. 15

Table 5. Drug classes detected by CARD's RGI in gut microbiome of COVID-19 patients' samples. 31

Table 6. Drug classes detected by CARD's RGI in water sewage samples. 32

Table 7. Drug classes detected by CARD's RGI in Osaka bay samples. 33

Table 8. Drug classes detected by CARD's RGI in skin AD patients' samples..... 34

Table 9. Drug classes detected by CARD's RGI in nose samples..... 36

Table 10. Drug classes detected by CARD's RGI in Tonga trench samples..... 36

Table 11. Drug classes detected by CARD's RGI in Mangrove samples 36

Table 12. Comparison between antiSMASH and deepBGC in terms of the detected BGCs' classes 39

Table 13. Potential application of some detected secondary metabolites 44

List of Figures

Figure 1. The distribution of filtered reads of all processed samples per each of the seven selected projects 10
Figure 2. Distribution of the assembled contigs per each project..... 11
Figure 3. The distribution of human abundance per each of the seven selected datasets. Excluded from downstream analysis 12
Figure 4. The Study workflow for the major steps of the pipeline used to screen the selected metagenomic & metatranscriptomic projects included in the dataset for BGCs..... 13
Figure 5. Barplot showing the relative abundance of all the detected microbial taxa, at the genus level, of the processed samples per each of the seven selected datasets. 17
Figure 6. PCA analysis for the ecosystems, based on the assigned microbiome taxa..... 18
Figure 7. t-SNE analysis for the datasets, based on the assigned microbiome taxa.....19
Figure 8. Distribution of the detected BGCs by antiSMASH in all datasets..... 20
Figure 9. Distribution of the absolute number of the detected BGCs classes by antiSMASH in all datasets 20
Figure 10. Distribution of the detected BGCs by deepBGC in all datasets. 24
Figure 11. Distribution of the absolute number of the detected BGCs classes by deepBGC in all datasets.. 25
Figure 12. Distribution the assigned product activities by deepBGC in all datasets ..26
Figure 13. Distribution the detected AMR genes families by CARD in all datasets ..29
Figure 14. Distribution the detected drug classes by CARD in all datasets 30
Figure 15. Distribution the detected resistance mechanisms by CARD in all datasets 30
Figure 16. Distribution of the number of contributed genus to BGCs with their respective contribution percentages as assigned by antiSMASH per dataset.....38
Figure 17. Distribution absolute number the different BGCs classes detected by antiSMASH per dataset 42
Figure 18. BGC hits detected by antiSMASH boxplot for each dataset, in relation to its assigned genus.....43
Figure 19. Heatmap for each dataset with each of the BGC hits from antiSMASH in relation to its assigned genus.....46
Figure 20. BGC hits detected by DeepBGC (product activity assigned) boxplot for each dataset, in relation to its assigned genus 48

Chapter 1: Literature Review & Study Objectives

Introduction

Global health challenges associated with antibiotic and chemotherapeutic resistances:

On 5 February 2018, the WHO summarized the global challenge associated with antibiotic resistance as follows, "antibiotic resistance is one of the major threats to global health, food security, and development, and its effect could extend to include everyone regardless of their ages or their country (Antibiotic resistance, 2020). A growing number of infections – such as pneumonia, tuberculosis, gonorrhoea, and salmonellosis – are becoming harder to treat as the antibiotics used to treat them become less effective". Moreover, these infections could lead to longer hospital stays, higher medical costs, and increased mortality (Antibiotic resistance, 2020). Every year, around 2 million people in the US are infected by a bacterial strain that is resistant to all existing antibiotics (Martens & Demain, 2017)**.** Furthermore, the resistance of chemotherapeutic anticancer drugs following therapy is a rising global health challenge (Holohan et al., 2013). Therefore, there is an unmet need and a great pressure on scientists and the health communities for discovering new alternative drugs to the current overused ones. Hence, exploring Natural Products (NPs) could provide a rich source of potentially effective drugs (Hernando-Amado et al., 2019). Deeper analysis of bacterial behavior in their respective communities is very crucial, recent studies shed the light on the key role of environments as a corner stone in not only the transmission of resistance genes between different bacterial species but also has an important role in emergence of pathogens with elevated level of resistance (Bengtsson-Palme et al., 2018).

Natural Products and their pharmaceutical importance

In nature, a wide range of secondary metabolites (SM) with different functional groups is produced by plants and microbes such as bacteria and fungi (Davies & Ryan, 2012). Unlike primary metabolites, SMs are not directly involved in vital processes (i.e. reproduction, growth & development) of the organism. However, these organic compounds which exist in many different chemical structures carry out a broad range of functions. In the mid-20th century, after the great discovery that some microbial natural products have an antimicrobial activity, an endless intensive research work has started and a wide range of microbial strains has been randomly screened for the

presence of natural byproducts with potential therapeutic activities (Davies & Ryan, 2012). These efforts yielded hundreds of thousands of SMs to be extracted and tested as antimicrobial agents (Davies & Ryan, 2012). Moreover, scientists harnessed the power of SMs to be utilized as antimicrobials, anticancer, immunosuppressant, and cholesterol-lowering agents and many others (Ruiz et al., 2010).

These natural products (NP) play a crucial role in drug discovery and development. According to David J. Newman in 2016, about 70% of anti-infective medicines originated from natural products (Newman & Cragg, 2016). Over 33 years, from 1981 to 2014, 32% of small molecule medicines approved by the FDA were natural products either unmodified (6%) or NP derivatives (26%) (Newman & Cragg, 2020). These drugs include different therapeutic classes such as antimicrobial, anticancer, diabetic, immunosuppressant, and cholesterol-lowering agents (Newman & Cragg, 2020).

Natural Products are encoded by Biosynthetic Gene Clusters (BGCs)

Previous reports investigating the characteristics of bioactive secondary metabolites revealed that the metabolic pathways of SMs are encoded by co-localized genes collectively called Biosynthetic Gene Clusters (BGCs) (Martin, 1992). Genes encoding for the biosynthetic pathway enzymes as well as their respective regulatory genes are contained in the BGC region (Keller et al., 2005). Notably, this fact paves the way for *in silico* mining of genomes and metagenomes for secondary metabolites through BGC neighborhood identification (Medema & Fischbach, 2015). So far, biosynthetic systems could be grouped into two major classes, Non-ribosomal peptide synthases (NRPS) and Polyketide synthases (PKS) (Weber & Kim, 2016). On the other hand, PKS and NRPS are responsible for synthesizing a wide and varied spectrum of bioactive natural products with much biomedical research and therapeutic applications such as antimicrobial, antifungal, and immunomodulatory agents, therefore PKS and NRPS are prevalent targets in genome mining for NPs (Ayuso-Sacido & Genilloud, 2005).

Tools for BGC mining and NP Discovery

NP mining: past, present and future

Prior to the omics era and the advancement of DNA sequencing technologies, exploring microorganisms for natural products was mainly conducted in the laboratory using culture-dependent techniques (Katz & Baltz, 2016). The classical way of natural products discovery typically consists of four main steps, starting with isolating the microbial samples, cultures enrichment, extracting candidate products and finally screening and screening their activities. One major drawback to this traditional method is the difficulty to culture microorganisms in the lab, besides that not all microbes can be grown in stable enrichments. To date, only a small fraction of microbial species could be cultured in the laboratory (Stewart, 2012). Growing microorganisms in the laboratory under diverse conditions was frequently used to produce and identify secondary metabolites without being able to specify their biosynthetic pathways at the genetic levels (Luo et al., 2014). Secondary metabolites functions and activities are usually characterized and validated through different biochemical assays. Recently, high throughput biochemical assays enabled the discovery of a wide range of unprecedented secondary metabolites with potential antimicrobial activities. One notable example, in a study of sugar fermentation in a Vibrio Cholerae culture, 49 out of 39,000 crude extracts screened were able to block fermentation pathways and 3 products with novel antimicrobial activities were identified representing a new class of broad-spectrum antibiotics (Chen et al., 2019). One major limitation linked to solely using biochemical assays to detect and characterize SMs is the fact that some SMs are formed at undetectable levels. Therefore, it will be more feasible to integrate biochemical assays with other approaches to capture a broader range of SMs produced in nature (Luo et al., 2014).

Omics Approaches in NP Discovery

Many pharmaceutical drugs which are approved for use by health authorities all over the world, have been discovered as a result of the traditional approaches of NP discovery. However, the rate of NP discovery has declined dramatically due to the difficulties of identifying novel compounds and the recurrent discovery of known compounds (Li & Vederas, 2009). Extraordinary opportunities for NP discovery through identification and characterization of biosynthetic gene clusters (BGCs) have

been created by genome sequencing technology (Jensen, 2016). While the early approaches in genetics were based on progressing from phenotype to genotype, the introduction of the next-generation sequencing techniques along with whole-genome sequencing approaches, creates databases waiting to be mined for novel BGCs discovery, characterization and synthesis through reverse genetic engineering approaches, from genotype to phenotype. The massive progress of genomic resources, especially microbial whole-genome sequencing, not only for the cultured organisms but also for the uncultured ones, has led to a notable paradigm shift in the uses of computational approaches in the discovery of bioactive natural products (Hannigan et al., 2019).

Genome mining is considered a highly time and cost effective approach in NP discovery because it allows researchers to examine huge genomic datasets whether it harbor biosynthetic gene clusters of interest or not, before undertaking any expensive and laborious biochemical steps to produce and extract the NP from microbial host. Omics approach makes it possible to identify a very large number of BGCs in different genomes and explore the chosen BGCs for experimental and systematic characterization (Chen et al., 2019).

In silico Tools for Biosynthetic Gene Clusters Identification

The rapid advances in the DNA sequencing techniques inspired the development of *in silico* tools and pipelines to mine microbial genomes and metagenomes for the presence of biosynthetic gene clusters (Table 1. showing a summary of the tools widely used to predict the biosynthetic gene clusters). The vast majority of them utilize Basic Local Alignment Search Tool (BLAST) or profile hidden Markov models (HMMs) searching tools as a base to identify the genetic signatures accountable for NP biosynthesis (Ren et al., 2020). These tools include NAPDOS, antiSMASH, NP.searcher and ClustScan, which are known for their high accuracy yet low levels of novelty. Moreover, genome mining can be leveraged by the presence of databases for the known BGCs such as antiSMASH (antibiotics & Secondary Metabolite Analysis Shell) and MIBiG (Minimum Information about a Biosynthetic Gene Cluster) (Blin et al., 2017) and (Starcevic et al., 2008). In 2019, Hannigan, Geoffrey D et al. introduced DeepBGC as a novel approach integrating deep machine learning with natural language

processing for a better outcome in terms of precision and accuracy in BGCs identification in microbial genomes (Hannigan et al., 2019).

Table 1. List of the widely used tools and pipelines for prediction of BGCs, data was modified from Ren H et al., 2020

Tools	Target(s)	Predicted BGC class	Reference
AntiSMASH	Bacteria & fungi	Wide range	Blin et al., 2013
NP.searcher	Bacteria	NRPS, PKS & NRPS/PKS	Li et al., 2009
ClustScan	Bacteria	NRPS & PKS	Starcevic et al., 2008
ClusterFinder	Bacteria	Wide range	Cimermancic et al.,2014
NaPDoS	Metagenomics	NRPS & PKS	Ziemert et al., 2012
eSNaPD	Bacteria	Wide range	Reddy et al., 2014
EvoMining	Bacteria	Wide range	Selem-Mojica et al., 2019
SMURF	Fungi	NRPS, PKS, NRPS/PKS & DMATS	Khaldi et al., 2010
PantiSMASH	Plant	Wide range	Kautsar et al., 2017
BAGEL	Bacteria	Bacteriocin & RIPP	Van Heel et al., 2013

Genomics and High-throughput Sequencing Technologies

Applying high throughput sequencing techniques in the study of microbial communities was the biggest reason behind creation of metagenomics research field; as it enables, for the first time, the study of different genomic sequences of co-existing microorganisms in a certain community (Ghurye et al., 2016). Sequencing technologies have been dramatically advanced during the past four decades. Sanger sequencing considered the first revolution discovery in modern genetic analysis because it allow complete genome sequencing for the first time. Later, genome sequencing became faster and much cheaper when the next-generation sequencing (NGS) technologies had appeared, regardless its advancement, NGS technologies have several drawbacks, most remarkably the problem of short reads (i.e. it produces up to several hundreds of base pairs). Recently, such pitfall has been solved by applying the third-generation sequencing technology which can produce long reads, up to several tens of kb, and genomic assemblies of extraordinary quality (van Dijk et al., 2018).

Long-read sequencing approaches had been enhanced over the recent years. Therefore, enabling the study of different genomic sequences and transcriptomes at an extraordinary resolution, therefore, metagenomics analysis could go deeper to the species level (Pootakham et al., 2017); (Kuleshov et al., 2016). In the near future, long-read sequencing has a great possibility to become a standard method in medical diagnosis. A recent SMRT study of a patient's genomic sequence showing undetected SV despite the aggressive genetic testing by other approaches (Merker et al., 2018).

Moreover, ambiguous regions in genomes are no longer big issues and can now be resolved, and more details will be elaborated from transcriptomes. Thus long-read methods are leading a series of revolutionary new discoveries in genomics research. Many long-read platforms are now available such as; PacBio, ONT and Illumina/10X Genomics SLR, table 2 summarize some of their strengths and highlighting their weaknesses (van Dijk et al., 2018).

Table 2. Strengths and weaknesses of some available long-read platforms, data was modified from van Dijk EL et al., 2018

Long-read platform	Strengths	Weaknesses
ONT	▪ Ultra long reads; a one Mb reads can be obtained ▪ Cost effective (e.g. MinION) ▪ Epigenetic modifications are directly detected ▪ Very fast library preparation ▪ Portable (e.g. SmidgION)	▪ High error rate ▪ Library preparation needs big amount of starting material ▪ Software versions subjected to numerous changes
PacBio	▪ High accuracy with CCS greater than 99% at 20 passes. ▪ Epigenetic modifications are directly detected ▪ Overcome repeats problem	▪ Expensive with high cost per Gb ▪ Library preparation needs big amount of starting material ▪ High error rate ▪ Only Sequel sequencer is available. ▪ Polymerase reactivity limit read length
Illumina/10X Genomics SLR	▪ High accuracy and low error rate ▪ Low cost per Gb ▪ No need for special equipment ▪ Library preparation needs small amount of starting material	▪ No real long reads ▪ Library preparation needs PCR amplification ▪ Epigenetic modifications couldn't detected directly ▪ Limited capacity (i.e. 384 wells)

The importance of searching for BGCs in metagenomes

Metagenomics is the study of genetic material of samples, recovered directly from the environment. Unlike the cultivated-based methods such as microbial genome sequencing, early environmental genomics rely upon sequencing of cloned specific genes (i.e. 16S rRNA gene) to generate a profile showing the microbial biodiversity in nature. As a result of applying metagenomics approaches, a whole world of endless different species has been discovered (Hugenholtz et al., 1998).

The vast majority of the discovered bioactive NP are products of microbial strains that can be cultivated in the laboratory. However, metagenomics studies open

unlimited possibilities to reach and study the uncultivated microbial communities which represent the bigger fraction of the underexplored microbial ecology. Furthermore, novel biosynthetic pathways are being discovered at higher rates compared to the old techniques of molecular biology. In addition, metagenomics would also serve as a great tool to study biocatalysts from the previously overlooked cultivated microbial strains which reflects a very good probability to discover novel compounds (Wilson & Piel, 2013).

AntiSMASH platform to detect BGCs

AntiSMASH is an inclusive *in silico* pipeline widely used to explore bacterial and fungal genome sequences to identify BGCs regions of a broad range of secondary metabolites (Medema & Fischbach, 2015) such as polyketides, terpenes, non-ribosomal peptides, bacteriocins, lantibiotics, siderophores, indolocarbazoles, aminocoumarins, aminoglycosides, beta-lactams, melanins, butyrolactones and others. Although, antiSMASH relies on signature gene profile HMMs for BGCs identification, they apply a greedy algorithmic method to extend the explored regions by 5, 10, or 20 kb on both sides hence closely localized clusters can be merged into what's called superclusters.

In the latest version (v.5.0) of antiSMASH pipeline (Blin et al., 2019), superclusters is relabeled as regions, and each region contains several mutually exclusive candidate BGCs for improved interpretation of hybrid clusters. Moreover, there are others additional options provided by antiSMASH such as, domain analysis and annotation of NRPS/PKS, core chemical structure prediction of non-ribosomal peptides and polyketides, comparative analysis of gene clusters by ClusterBlast, and protein family analysis of secondary metabolites (smCOG). The output can easily be visualized through an interactive XHTML page with a user-friendly interface (Ren et al., 2020). It is worth mentioning that there is another derivative of antiSMASH used for plant genome mining called plantiSMASH and it has ability to identify biosynthetic pathways between and within gene clusters by co-expression analysis, and also it can be used to study the evolutionary conservation of each gene cluster through comparative genomic analysis (Kautsar et al., 2017).

Deep learning method to detect BGCs

While they considered the gold standard for genome mining, current available pipelines, such as ClusterFinder and antiSMASH, are based mainly on signature gene profile HMMs for BGCs identification but they miss the ability to remember the effects of position dependencies between distant units or order information (Yoon, 2009); (Eddy, 2004). This leads to the fact that such tools, HMM-based, could not grasp higher order information among units (Yoon, 2009; Eddy, 2004), as a result they had a limited ability to detect BGCs.

To address this algorithmic limitation, Hannigan, Geoffrey D et al. implement a deep learning approach, DeepBGC, as a novel pipeline integrating deep machine learning with natural language processing (NLP) for a better outcome in terms of precision and accuracy in BGCs identification in microbial genomes (Hannigan et al., 2019). To overcome limitation of HMM-based tools, DeepBGC applying both Recurrent Neural Networks (RNNs) and vector representations of protein family (Pfam) domains (Finn et al., 2016) which together have the ability of inherently sensing short and long term effects of position dependency between neighboring and distant entities (Sepp Hochreiter et al., 2007).

DeepBGC applies a Bidirectional Long Short-Term Memory (BiLSTM) RNN besides a word embedding skip-gram neural network, word2veclike, called pfam2vec (S. Hochreiter & Schmidhuber, 1997). Implementation of DeepBGC produce a higher performance compared to the leading algorithms in terms of the accuracy of BGC detection from different genome sequences and the ability of identification of novel classes of BGCs. Additionally, DeepBGC can classify the identified BGCs based on their corresponding product classes and the product molecular activity by using a generic random forest classifiers.

DeepBGC considered a new powerful tool which when applied to bacterial reference genomes could identify biosynthetic gene clusters coding for bioactive molecules with putative antimicrobial activity that never identified by the other existing pipelines. Moreover, the power of this tool might be used in metagenomic analyses in addition to microbial reference genome, this might leads to a new era of improved BGC detection and unlimited possibilities to identify novel BGCs (Hannigan et al., 2019).

Study Objectives

In this study, selected samples pertaining to seven selected metatranscriptomic / metagenomic projects were subjected to assembly, taxonomic assignment to the reads and assembled contigs. Aim of this study is two-fold. Firstly, the assembled contigs were then investigated by two major distinct computational methods, namely antibiotics and Secondary Metabolite Analysis Shell (antiSMASH) and deep-learning (deepBGC) methods. A comparative study was performed to determine the biosynthetic gene clusters (BGCs) present in each of the included samples, as well as comparing their taxonomic differences. Secondly, the assembled contigs were also analyzed to determine the antimicrobial resistance (AMR) genes present in each samples by using Resistance Gene Identifier (RGI) algorithm which is a part of the Comprehensive Antibiotic Resistance Database (CARD).

Chapter 2: Materials & Methods

Samples and Assembly

Whole metagenome samples were obtained from NCBI Sequence Read Archive (SRA) using prefetch then the downloaded SRA files were converted into paired-ended FASTQ using fastq-dump. FASTQ files were processed for quality control to remove adaptor sequences, trim low-quality ends, and remove short reads using fastp (Chen et al. 2018). Filtered sequences were sub-sampled to one million reads per sample (run) using Seqtk https://github.com/lh3/seqtk, Figure 1 showing the distribution of filtered reads of all processed samples per each of the seven selected projects. Sequence reads were assembled using MEGAHIT (Li et al. 2015), Figure 2 showing the distribution of the assembled contigs per each project. Assembled contigs were taxonomically classified using Kraken 2 (Wood et al. 2019). Contigs were filtered for a minimum size of 1,000 nucleotides.

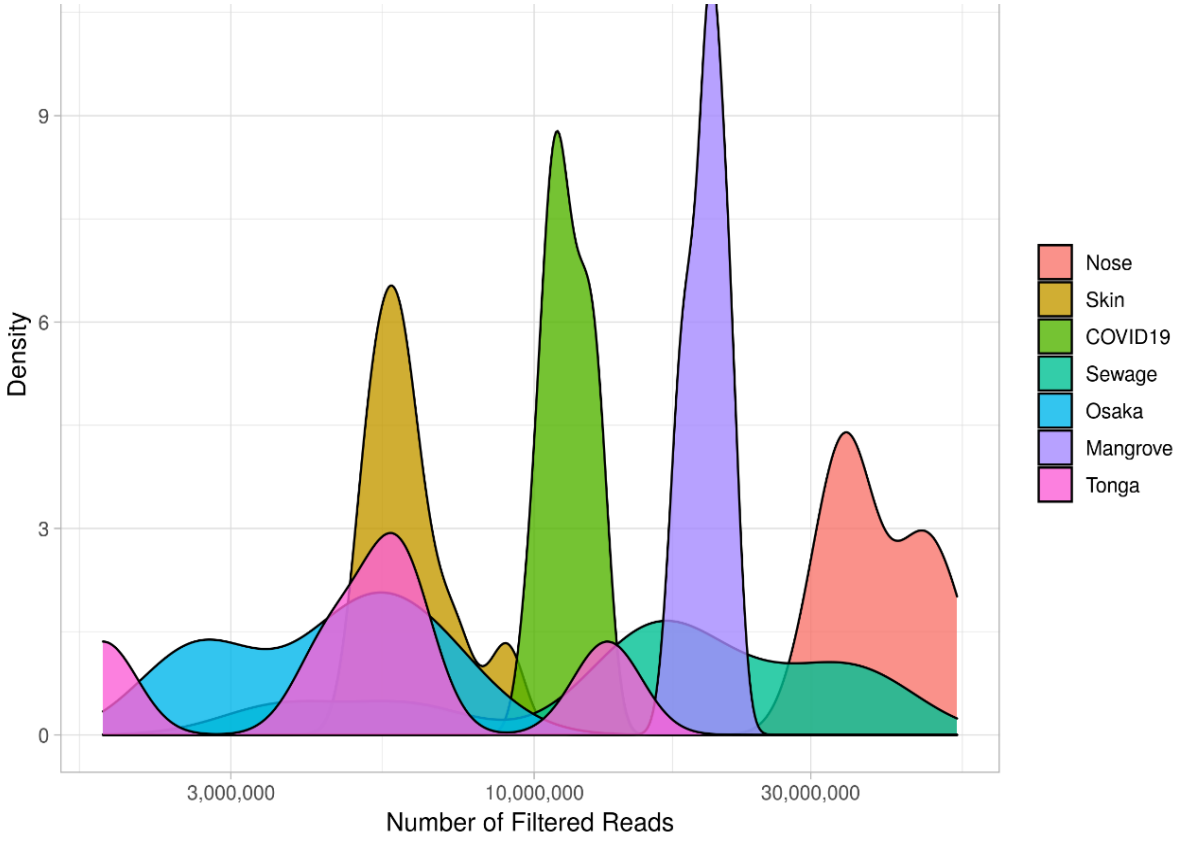

Figure 1. The distribution of filtered reads of all processed samples per each of the seven selected projects.

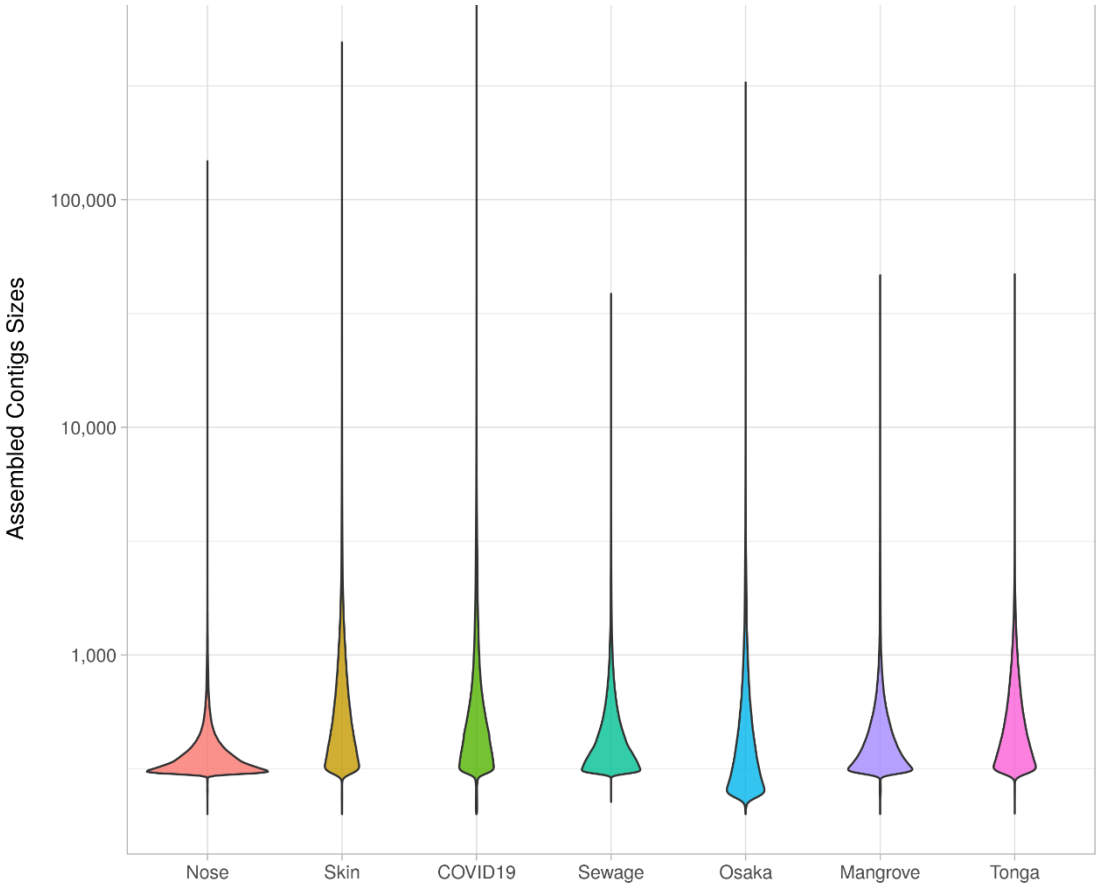

Figure 2. Distribution of the assembled contigs per each project

Taxonomic analysis, annotation and bioinformatic visualization

Samples were taxonomically classified on the sequence reads using Kraken 2 (Wood et al. 2019) with the default taxonomy database. The abundance of the different taxonomic levels (species, genus, family, etc.) was estimated using Bracken (Lu et al. 2017).

Using deepBGC tool for BGC mining

Biosynthetic gene clusters were predicted in the assembled contigs using deepBGC (Hannigan et al. 2019). Contigs classified as human were excluded from downstream steps (Figure 3).

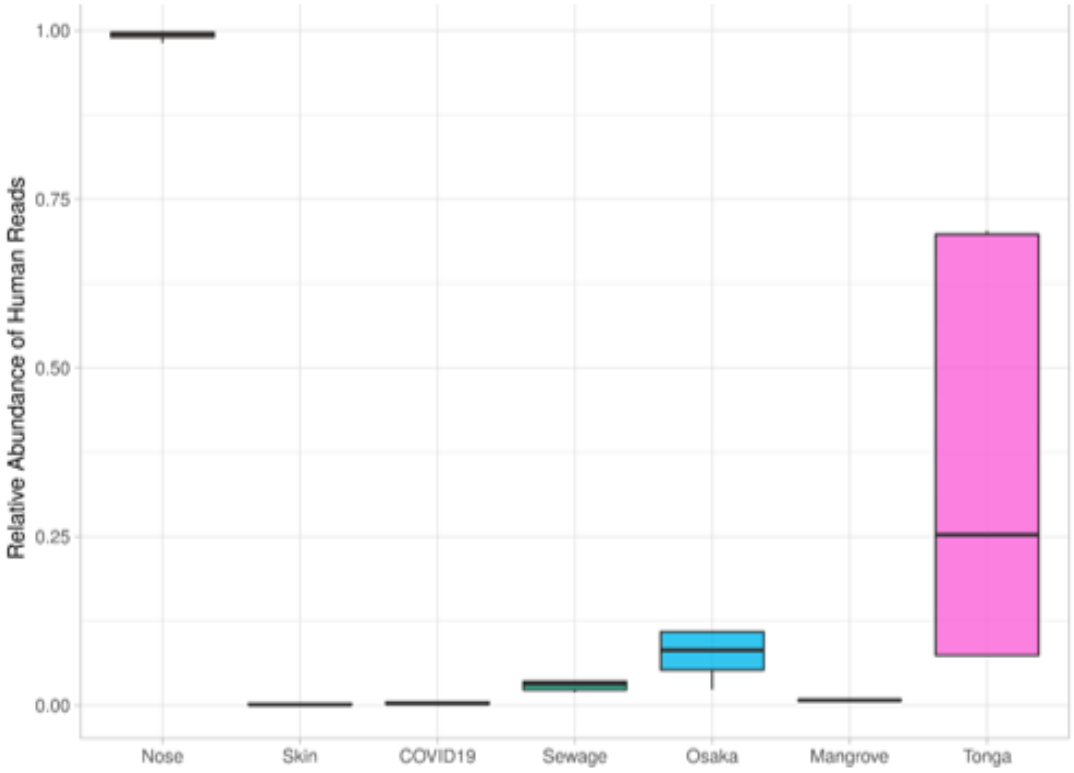

Figure 3. The distribution of human abundance per each of the seven selected datasets. Excluded from downstream analysis

Using antiSMASH tool for BGC mining

The antibiotics and secondary metabolite analysis shell (antiSMASH) platform was utilized for detection of BGCs. antiSMASH bacterial version 5.0 was used with default parameters (Medema et al. 2011).

Using CARD's RGI algorithm for AMR genes detection

The Resistance Gene Identifier (RGI) algorithm present in the Comprehensive Antibiotic Resistance Database (CARD) was exploited for determination of AMR genes, drug classes and their resistance mechanisms. The following RGI criterions for detection were applied, perfect, strict & loose, partial genes included, 95% identity nudge used and low quality coverage was used in the sequence quality option. (Alcock BP, Raphenya AR, Lau TTY, et al. 2020). Loose hits were excluded from downstream steps.

Statistical Analyses

Analytical and visualization analyses were performed using R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Chapter 3: Results

Assembly pipeline

The study workflow is illustrated in Figure 4, and the details of samples used in this analysis are available in Table 3. In this study, a total number of 65 samples from seven selected projects were processed using both antiSMASH and deepBGC pipelines for BGCs mining and CARD's RGI algorithm for AMR genes detection. The total number of reads used was 1,139,543,039 yielded 1,100,630,009 filtered reads and generating a total of 4,325,515 contigs (Table 4). The contigs (assembled metagenomes and metatranscriptomes) from the seven selected projects included in the dataset were investigated by two major distinct computational methods (i.e. antiSMASH and deepBGC) in addition to CARD's RGI algorithm for BGCs mining and AMR genes detection, respectively, the workflow is depicted in Figure 4, and a comparative study was performed to determine the BGCs present in each of the included samples along with detection of AMR genes with their corresponding mechanisms of action and drug classes which they were confer resistance to it. The assembly metrics are denoted in Table 4.

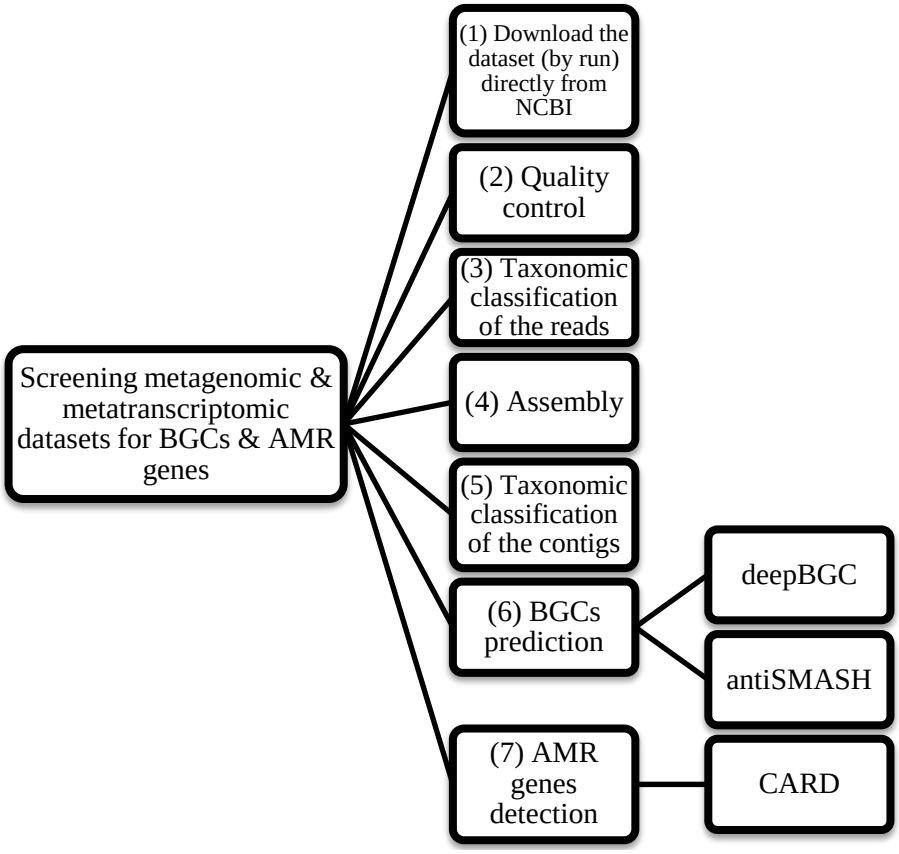

Figure 4. The Study workflow for the major steps of the pipeline used to screen the selected metagenomic & metatranscriptomic projects included in the dataset for BGCs

Table 3. The selected metagenomic and metatranscriptomic projects included in the dataset. The accession numbers of the seven selected datasets along with their corresponding abbreviated names and number of processed samples are denoted. A total number of 65 samples were processed using both pipelines, antiSMASH and deepBGC for BGCs mining and CAR's RGI for AMR genes detection.

#	Accession number	Dataset Name (Abbreviated)	Number of processed samples per dataset
1	PRJDB6156	Osaka	9
2	PRJNA340165	Tonga	5
3	PRJNA472006	Nose	12
4	PRJNA489681	Skin AD	10
5	PRJNA624223	COVID-19	10
6	PRJNA629394	Mangrove	10
7	PRJEB13831	Sewage	9
		Total number of samples	**65**

Table 4. Assembly metrics denoted for each sample from the selected metagenomic and metatranscriptomic projects included in the dataset.

Project	Accession	Organism	Organism Groups	Data type	Location	Reg. Date	Sample ID	# of Bases	# of Reads	# of Filtered Reads	# of Contigs	Av. Contig Size (bp)	Largest Contig Size (bp)
Sequencing of Osaka Bay see water metatranscriptome (at 5 m depth)	PRJDB6156	Marine metagenome	Metagenomes; ecological metagenomes	Transcriptome or gene expression	Japan: Osaka Bay	2019-08-01	DRR099940	387.8 M	2,667,479	2,649,580	7,657	487	9,321
							DRR099941	1.1 G	7,509,742	7,382,595	45,800	559	59,858
							DRR099942	749.7 M	5,334,601	5,239,370	26,283	523	18,690
							DRR099943	864.7 M	5,921,570	5,813,994	36,796	454	139,294
							DRR099944	932.3 M	6,505,903	6,377,636	14,877	945	327,531
							DRR099945	382.6 M	2,612,159	2,581,855	15,734	737	21,593
							DRR099946	397.9 M	2,769,912	2,719,784	8,059	1,216	37,754
							DRR099947	673.22 M	4,642,688	4,581,593	18,774	378	8,907
							DRR099948	663.4 M	4,650,613	4,593,213	13,583	846	48,051
							TOTAL (M)	**6151.62**					
Investigation of the metagenome of the Tonga trench sediment (at 9.2 km water depth and up to 2 m sediment depth)	PRJNA340165	Marine sediment metagenome	Metagenomes; ecological metagenomes	Metagenome	Pacific Ocean	2016-08-25	SRR4069403	417.20 M	1,846,010	1,804,893	23,966	481	16,294
							SRR4069404	1.30 G	5,826,299	5,732,762	110,986	572	47,068
							SRR4069405	1.03 G	4,564,804	4,443,188	3,675	519	2,701
							SRR4069406	3.09 G	13,747,868	13,377,953	4,220	591	4,378
							SRR4069408	1.34 G	5,974,450	5,868,302	927	375	5,306
							TOTAL (M)	**7177.20**					
Comparative metagenomic analysis to assess the relationship between human skin microbiota stability and patients with atopic dermatitis (49 subjects, 33 AD patients and 16 healthy controls)	PRJNA489681	Multiple	Metagenomes	Metagenome	Singapore	2018-09-06	SRR7802475	1.87 G	9,346,770	8,963,662	29,068	1,134	327,256
							SRR7802341	1.49 G	7,438,112	7,363,758	43,096	627	388,652
							SRR7802306	1.37 G	6,854,660	6,577,935	81,181	687	160,560
							SRR7802339	1.09 G	5,456,812	5,402,061	23,516	677	312,511
							SRR7802351	1.22 G	6,112,223	5,715,559	29,534	1,386	491,444
							SRR7802349	1.03 G	5,217,253	5,170,642	39,712	811	340,940
							SRR7802476	1.27 G	6,365,734	6,140,000	27,564	943	104,289
							SRR7802335	1.19 G	5,973,209	5,740,124	56,361	673	275,594
							SRR7802352	1.02 G	5,143,013	5,100,160	128,979	815	129,231
							SRR7802406	1.22 G	6,111,603	5,903,453	65,083	712	145,141
							TOTAL (M)	**12770**					
Metagenomic data from Mangrove sediment microbiome along South China (samples are from Aegiceras corniculatum soil)	PRJNA629394	Sediment metagenome	Metagenomes; ecological metagenomes	Metagenome	South China	2020-04-29	SRR11734720	3.64 G	18,209,386	17,427,535	9,660	422	5,337
							SRR11734598	3.83 G	19,129,190	18,342,777	10,852	424	7,412
							SRR11734613	3.84 G	19,180,698	18,338,146	25,121	461	43,105
							SRR11734656	4.15 G	20,768,796	20,169,382	52,455	465	32,405
							SRR11734640	4.16 G	20,805,669	20,243,850	48,897	516	46,618
							SRR11734616	4.12 G	20,591,363	19,765,013	29,634	459	45,747
							SRR11734716	4.11 G	20,559,523	19,722,052	13,382	423	10,080
							SRR11734719	4.34 G	21,676,963	20,830,425	15,150	426	7,083
							SRR11734596	4.53 G	22,671,952	21,791,702	17,194	426	6,779
							SRR11734597	4.52 G	22,602,093	21,705,388	16,722	424	6,841
							TOTAL (M)	**41240**					

Gut microbiome alterations and longitudinal kinetics in 15 COVID-19 patients. Tao Zuo et al., 2020	PRJNA624223	Feces metagenome	Metagenomes; organismal metagenomes	Raw sequence reads	Hong Kong	2020-04-10	SRR12328926	2.85 G	10,119,197	9,893,607	15,215	1,195	732,744
							SRR12328948	3.10 G	10,831,211	10,698,264	7,994	3,351	403,502
							SRR12328907	3.05 G	10,776,976	10,650,748	156,679	1,299	524,513
							SRR12328910	3.16 G	11,141,802	10,735,634	409,228	490	385,138
							SRR12328904	3.32 G	11,404,379	11,294,382	136,136	1,363	472,741
							SRR12328942	3.20 G	11,471,303	11,240,588	31,516	1,614	585,648
							SRR12328943	3.41 G	12,238,736	12,012,970	14,049	1,123	349,596
							SRR12328897	3.52 G	12,313,208	12,205,473	149,575	1,253	481,689
							SRR12328903	3.68 G	13,018,722	12,883,352	116,637	1,560	493,077
							SRR12328951	3.76 G	13,019,613	12,900,042	25,191	2,502	411,905
							TOTAL (M)	**33050**					
Human skin metagenome and 16S (Epithelium of external nose)	PRJNA472006	Human skin metagenome	Metagenomes; organismal metagenomes	Raw sequence reads	Denmark: Copenhagen	2018-05-18	SRR9696273	8.2G	32,489,498	31,778,213	94,964	358	37,829
							SRR9696274	13.9G	55,186,070	53,638,515	43,565	337	16,576
							SRR9696275	9.1G	36,087,682	35,391,282	18,676	335	5,245
							SRR9696276	9.1G	35,935,443	35,178,043	30,113	335	16,689
							SRR9696277	8.8G	34,846,139	34,191,678	10,537	376	16,683
							SRR9696278	13.2G	52,223,806	50,934,770	44,869	337	4,667
							SRR9696279	9.6G	37,987,223	37,080,691	44,017	332	16,668
							SRR9696280	8.9G	35,189,492	34,402,185	3,275	329	3,917
							SRR9696281	12G	47,577,639	46,427,813	45,784	339	16,690
							SRR9696282	11.2G	44,434,656	43,402,757	42,952	338	16,668
							SRR9696283	7.8G	30,778,753	30,118,478	7,434	513	147,689
							SRR9696284	11.9G	47,163,030	46,080,142	23,316	331	16,668
							TOTAL (M)	**123700**					
Global surveillance of infectious diseases and antimicrobial resistance from sewage	PRJEB13831	Sewage	Metagenomes	Raw sequence reads	Global project	2019-02-01	ERR1713410	8.2G	27,296,861	25,450,270	253,567	396	23,682
							ERR1713411	5.8G	19,360,109	18,017,557	233,836	394	32,340
							ERR1726031	1.2G	3,919,046	3,659,867	59,188	382	6,765
							ERR1726032	4.6G	15,288,191	14,959,449	222,760	396	31,300
							ERR1726033	11.5G	38,069,060	36,914,867	236,108	397	27,547
							ERR1726034	1.9G	6,416,235	6,078,322	107,287	386	15,425
							ERR1726035	4.8G	15,847,330	14,650,560	195,063	392	14,190
							ERR2592282	5.8G	19,303,711	18,088,371	232,720	393	38,612
							ERR2592343	13G	43,018,798	36,090,777	218,766	391	12,778
							TOTAL (M)	**56800**					
TOTAL							**65 Samples**	**280888.82**	**1,139,543,039**	**1,100,630,009**	**4,325,515**	**-**	**-**

Taxonomical Assignment

To understand the dynamics of SM production in the different environments, this require to get the taxonomical assignment for the sequence reads of each sample. Samples were taxonomically classified on the sequence reads using Kraken 2 with the default taxonomy database. The abundance of the different taxonomic levels (species, genus, family, etc.) was estimated using Bracken. This exercise resulted into understanding the community structure and the abundance of each microbial group within each sample under test. Figure 5 showed how the relative abundance at genus level differ between samples of each projects, and it was clear that samples of some projects were dominated by few signature genera such as the sample of Osaka which were dominated mainly by *Pseudomonas* and *Synechococcus*. On the other hand, metagenomic skin samples of AD patients were dominated mainly by *Cutibacterium*, while *Corynebacterium* appeared like it stands alone in the samples of the Human skin metagenome from epithelium of external nose project (Nose).

Figure 5. Barplot showing the relative abundance of all the detected microbial taxa, at the genus level, of the processed samples per each of the seven selected datasets.

To show how the different samples will be clustered based on the relative abundance of taxa we constructed both a PCA and t-SNE graphs (Figures 6 & 7). Projects like Osaka bay and water sewage appeared completely separated from the other five projects, while the rest shows some connections which mainly due to the presence of common genera, such as *Corynebacterium* and *Cutibacterium* which explained the presence of some samples from the Human skin metagenome from epithelium of external nose and Tonga trench project around the samples of the skin project of AD patients. These findings might be of great impact on understanding the dynamics of SM production in different environments.

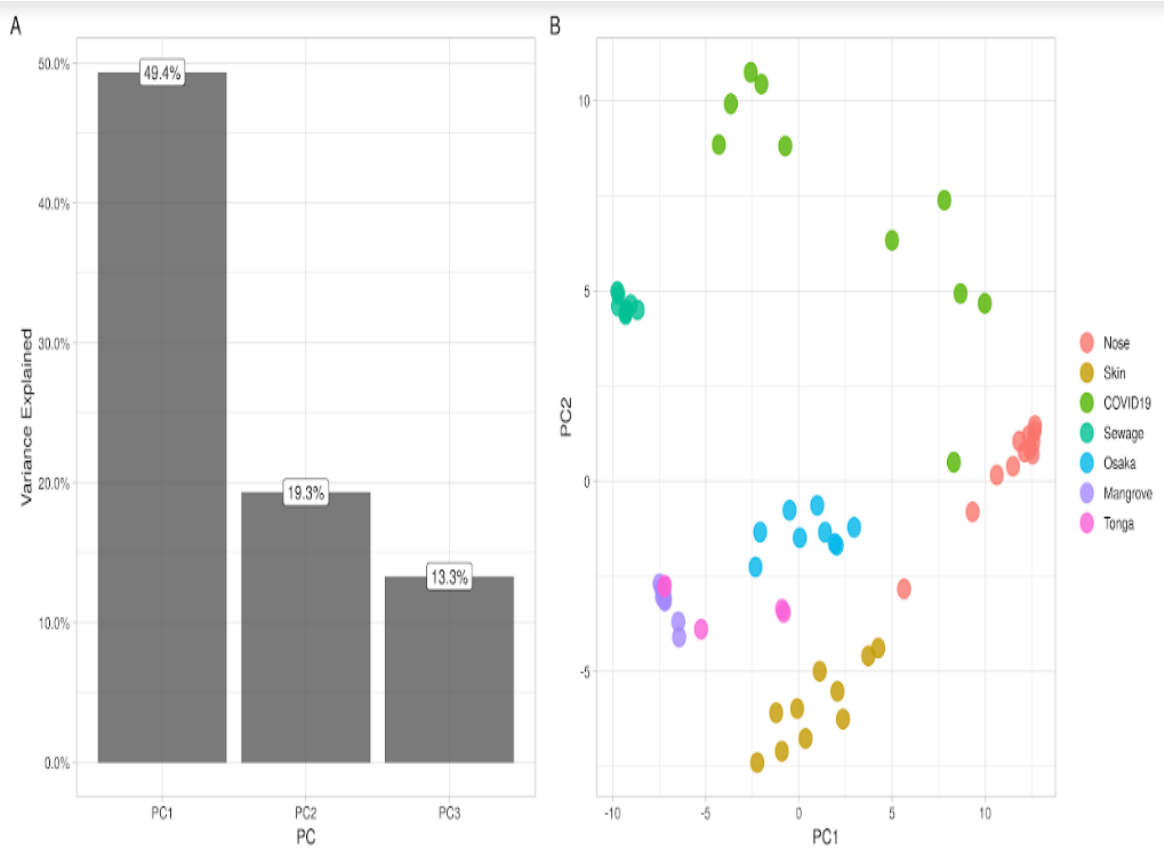

Figure 6. PCA analysis for the ecosystems, based on the assigned microbiome taxa. (A) Plot showing the most significant Principal Components, PC1, PC2 & PC3, all together represent 82% variations. (B) PCA biplot of the different samples from each of the seven selected datasets. PC1 & PC2 representing the most significant principle components and they cumulatively represent 68.7% and different samples were clustered separately based on their relative abundance of the taxa. Different datasets were color coded.

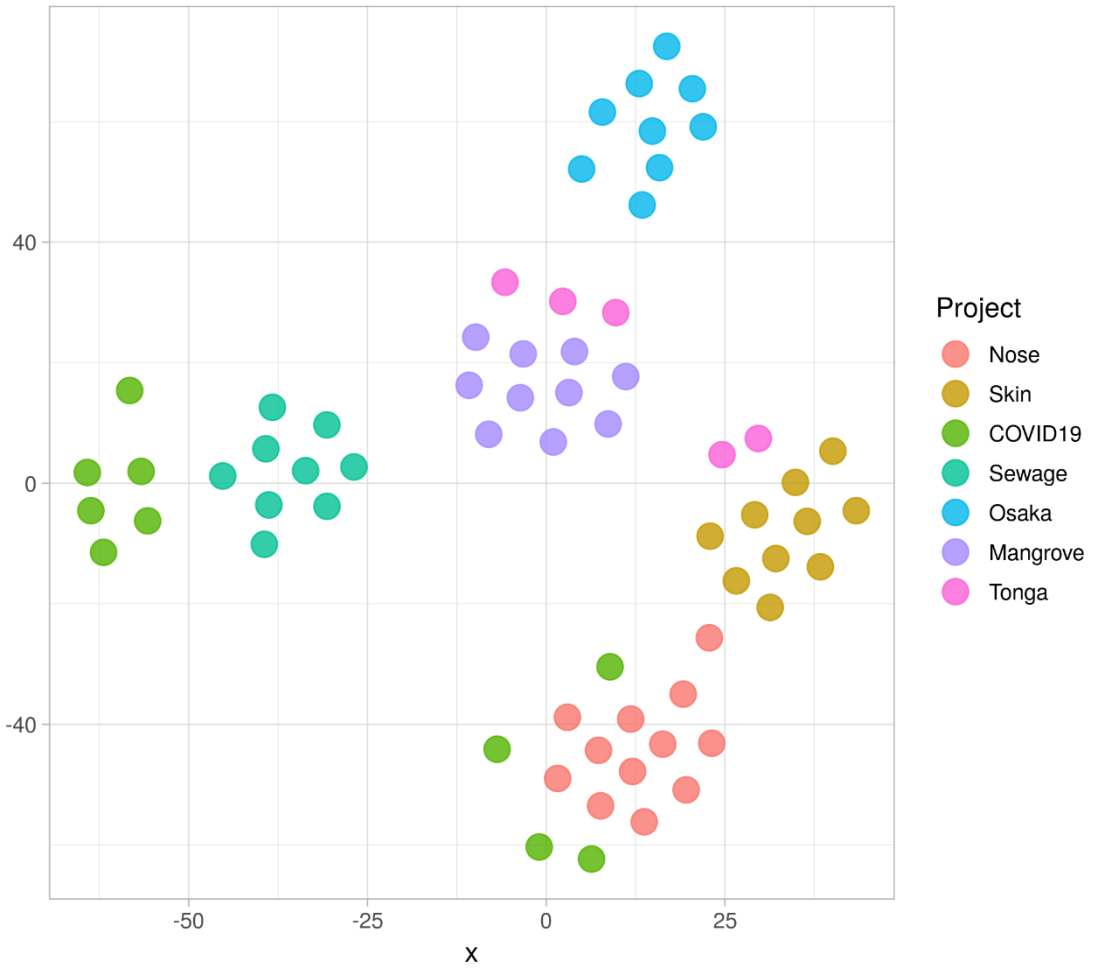

Figure 7. t-SNE analysis for the datasets, based on the assigned microbiome taxa.

BGC profile of samples in the dataset as detected by antiSMASH

By using antiSMASH a total 776 BGCs regions were detected from the selected projects, 45 samples from 65 processed samples gave hits with antiSMASH (Figure 8). 26 different BGC classes were detected (Figure 9), the seven major detected classes which collectively represent about 80% of the total detected BGCs classes were NRPS (23%), bacteriocin (15%), NRPS-like (10%), terpene (10%), sactipeptide (9%), arylpolyene (7%) and siderophore (5%). About 35% of the detected BGCs came from 5 different bacterial genera, *Pseudomonas* contributed the most with 12% and it was obvious that the most dominant species was *Pseudomonas sp. J380* which contributed alone by 10% of the total percentage of the detected BGCs. The genus *Gordonia* came in the second place with 8%, while the genus Corynebacterium produced 7% of the detected BGCs and both *Cutibacterium* and *Blautia* genera contributed by the same percentage of the detected BGCs, about 4% for each genus. To figure out the major differences between the processed samples from each environments in terms of BGCs

contents and their corresponding microbial contributors, we deeply analyzed the processed samples from each dataset and the results were explained in details here-under.

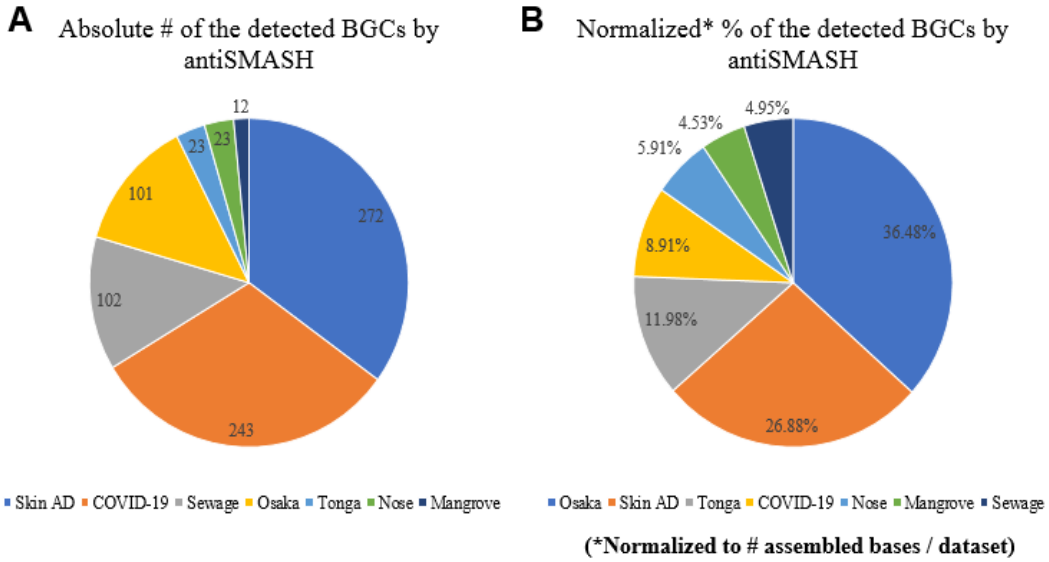

Figure 8. Distribution of the detected BGCs by antiSMASH in all datasets (A) Distribution of the absolute number of the detected BGCs by antiSMASH (B) Distribution of the normalized percentages of the detected BGCs by antiSMASH. Percentages were normalized to the number of assembled bases per dataset.

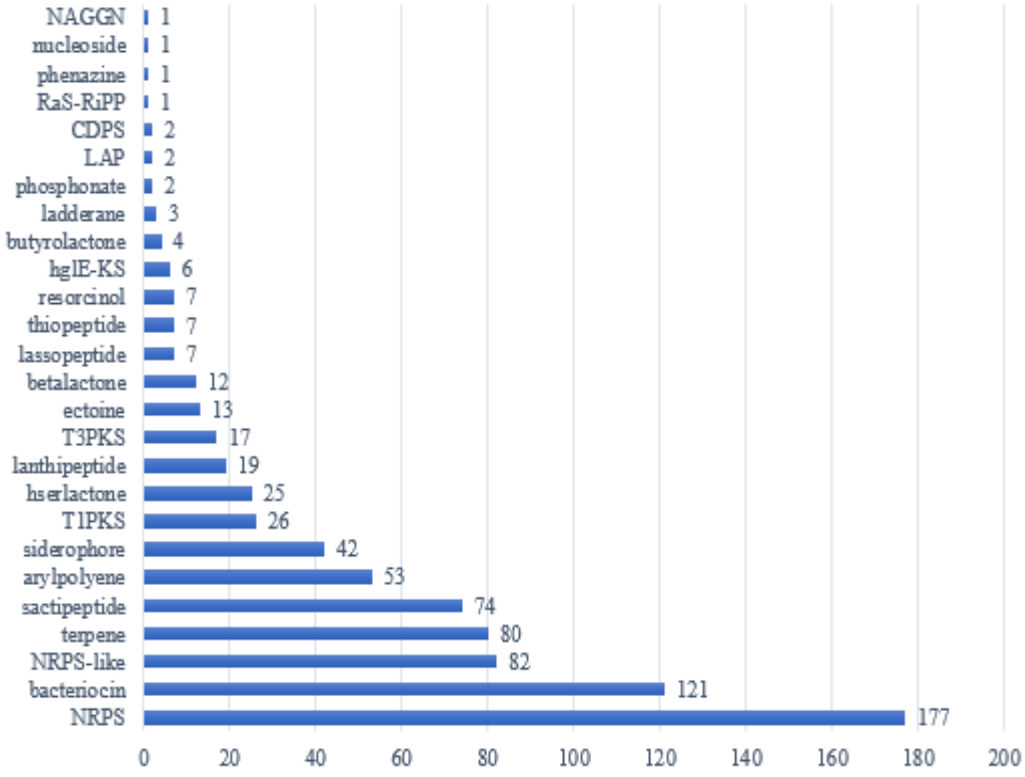

Figure 9. Distribution of the absolute number of the detected BGCs classes by antiSMASH in all datasets.

About 36.48%* (i.e. 101 BGCs) of the detected BGCs of 9 different classes were from the metatranscriptomic marine samples from Osaka Bay see water project. Only 6 samples out of 9 processed samples gave hits with antiSMASH pipeline. NRPS and bacteriocin represent the vast majority of the detected classes by 64%, NRPS class came in the first place by 44% while the percentage of the detected bacteriocin class was 20% and the percentages of the rest seven detected classes were as follows, 9% terpene, 7% NRPS-like, 7% arylpolyene, 5% hserlactone, 5% siderophore, 3% betalactone and only 1 BGC was NAGGN. 95% of the detected BGCs was produced by two genera, *Pseudomonas* and *Synechococcus*. *Pseudomonas* contributed the most with 82% and it was obvious that the most dominant species was *Pseudomonas sp. J380* which contributed alone by 74% from the total percentage of the detected BGCs. *Synechococcus* came in the second place and contributed by 13% of the total detected classes. * Percentage was normalized to the total number of bases.

The results of metagenomic samples from the gut microbiome project of COVID-19 patients were analyzed. In this study, 10 samples were processed, all samples gave hits and antiSMASH detected 16 different BGCs classes with a total number of 243 BGCs. Only 5 classes out of 16 different classes, represent about 80% of the total detected classes as follows; 28% was sactipeptide, NRPS represent 19%, bacteriocin was 16%, arylpolyene and lanthipeptide represent 12% and 7% respectively. The rest of detected classes (11 classes) which collectively represent about 19% were, NRPS-like, terpene, T3PKS, lassopeptide, betalactone, resorcinol, siderophore, thiopeptide, nucleoside, butyrolactones and ladderane. About 43% of the detected BGCs were from 5 genera, 13% of BGCs was produced by *Blautia*, and 12% was from *Lachnospiraceae*, 7% was *Bacteroides*, 6% was *Faecalibacterium* and *Streptococcus* contributed with 5% of the detected BGCs.

Ten skin microbiome metagenomic samples were randomly chosen from a comparative metagenomic analysis study conducted in Singapore to assess the relationship between human skin microbiota stability and patients with atopic dermatitis. All samples gave hits with antiSMASH and it detected 19 different BGCs classes with a total number of 272 BGCs. Out of the 19 detected classes, 6 represent about 81% of the total number of the detected BGCs. These classes were NRPS, NRPS-like, siderophore, terpene, bacteriocin and T1PKS which represent 27%, 17%, 10%,

10%, 9% and 8% respectively. The rest of detected classes (13 classes) which collectively represent about 19% were, ectoine, T3PKS, hserlactone, thiopeptide, betalactone, lanthipeptide, arylpolyene, CDPS, LAP, hglE-KS, ladderane, butyrolactones and lassopeptide. More than half of the detected BGCs (i.e. about 57%) were produced by 4 genera; *Gordonia*, *Corynebacterium*, *Cutibacterium* and *Staphylococcus* which represent 24%, 15%, 12% and 6% of the total detected BGCs respectively.

Five marine sediment metagenomic samples from Tonga trench sediment in the Pacific Ocean were processed by antiSMASH to screen for BGCs contents. Only 2 samples gave hits and antiSMASH detected a total of 23 BGCs with 9 different classes. The detected classes with their corresponding percentages were as follows, (17%) arylpolyene, (17%) NRPS-like, (17%) hglE-KS, (13%) bacteriocin, (9%) phosphonate, (9%) terpene, (9%) hserlactone, (4%) T1PKS and (4%) NRPS. There was no much data about the microbial composition of the processed samples because antiSMASH could not assign about 43% of the detected BGCs to any microbial species.

Another ten metagenomic samples from Mangrove sediment microbiome along South China were processed and 5 of them gave hits with antiSMASH. A total of 12 BGCs with 6 different classes were detected. The detected classes with their corresponding percentages were as follows, (33%) bacteriocin, (17%) arylpolyene, (17%) NRPS-like, (17%) terpene, (8%) lassopeptide and (8%) NRPS. There was no much data about the microbial composition of the processed samples because antiSMASH could not assign about 58% of the detected BGCs to any microbial species.

Twelve samples were a Human skin metagenome from epithelium of external nose from a study conducted in Copenhagen; Denmark. Only 3 samples gave hits with antiSMASH and it was obvious that only one dominant genus, Corynebacterium produced 5 different classes of BGCs and a total 23 BGCs were detected as follows; 9 NRPS representing 39%, 5 terpene (22%), 5 siderophore (22%), 2 NRPS-like (9%) and 2 T1PKS (9%). There was no much data about the microbial composition of the processed samples because antiSMASH could not assign about 53% of the detected BGCs to any microbial species.

The last 9 samples were from the Global Sewage Project. A total of 102 BGCs with 12 different classes were detected. The detected BGCs classes with their corresponding percentages were as follows, (30%) bacteriocin, (28%) terpene, (10%) hserlactone, (8%) NRPS-like, (7%) arylpolyene, (5%) sactipeptide, (3%) resorcinol, (3%) T3PKS, (2%) ectoine, (2%) butyrolactones, (1%) phenazine and (1%) RaS-RiPP. Almost 50% of the produced BGCs was from 3 major genera; *Streptococcus* comes in the first place with 27%, *Neisseria* produced about 15% and 6% was from *Polaromonas*. What is interesting about this project is that there are few unique BGCs classes detected by antiSMASH which are not appear in the previous 6 projects, such as resorcinol, ectoine, phenazine and RaS-RiPP. Moreover, these classes are not produced by the dominant genera (i.e. resorcinol is produced either by *Brevundimonas* or *Pasteurellaceae*, ectoine produced by *Arcobacter* and phenazine was produced by *Escherichia coli*) each genera represent only about 1% of the total microbial community. We discussed this point with some details elsewhere in this study.

BGC profile of samples in the dataset as detected by DeepBGC

Before we decide to use deepBGC pipeline, we did a pilot trial to test deepBGC output and the results were very interesting and rich with huge amount of data compared to antiSMASH. Although it is not an objective of this study to perform a comparative analysis between the results obtained by both pipelines, however, we decide to use deepBGC to get a deeper insight and grasp more information about the processed samples and their corresponding communities.

For a better understanding of deepBGC results here are some important points about this algorithm. The current deepBGC pipeline could detects only six different BGCs classes, five specific (i.e. Polyketide, NRP, RiPP, Saccharide and Terpene) and one unspecific annotated by the algorithm as “other”. In addition to the huge amount of data generated by deepBGC there is another major advantage of it because it could assign four different products’ activities (i.e. antibacterial, antifungal, inhibitor and cytotoxic) to each detected BGCs with a very high coverage percentage, almost 96% of processed samples. We tried to figure out the annotation mechanism of deepBGC, we discovered that there is a scoring system for both class and product activity annotations, the pipeline assigned class and activity for hits with scores ≥ 0.5 and if there are more

than one class or activity with a score ≥ 0.5, all will be annotated to the same hit, separated by a hyphen sign (-).

Unlike antiSMASH, deepBGC detected large number of BGCs classes and a total 79,771 BGCs were detected from the selected datasets (Figure 10), moreover all the 65 samples gave hits. DeepBGC assigned BGCs classes to around 20% of hits (15,714 hits) as follows; 39% of hits was Polyketide, 18% RiPP, 18% Saccharide, 10% others, 7% NRP, 4% Terpene, 2% Polyketide-Terpene, 1% NRP-Polyketide and 1% Saccharide-Terpene (Figure 11).

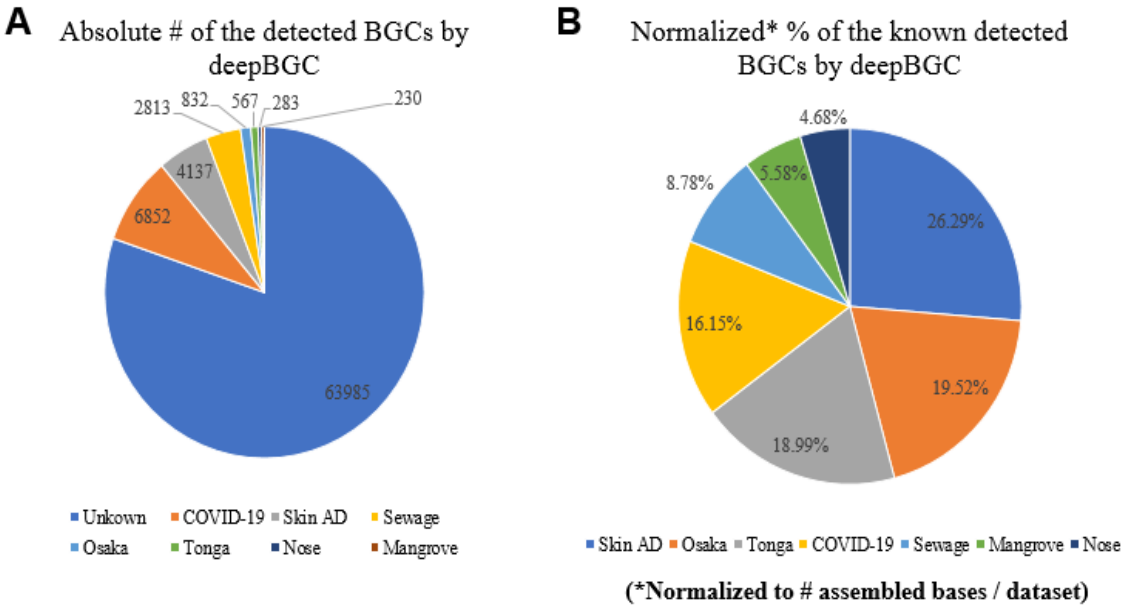

Figure 10. Distribution of the detected BGCs by deepBGC in all datasets (A) Distribution of the absolute number of the detected BGCs by deepBGC (B) Distribution of the normalized percentages of the detected BGCs by deepBGC. Percentages were normalized to the number of assembled bases per dataset.

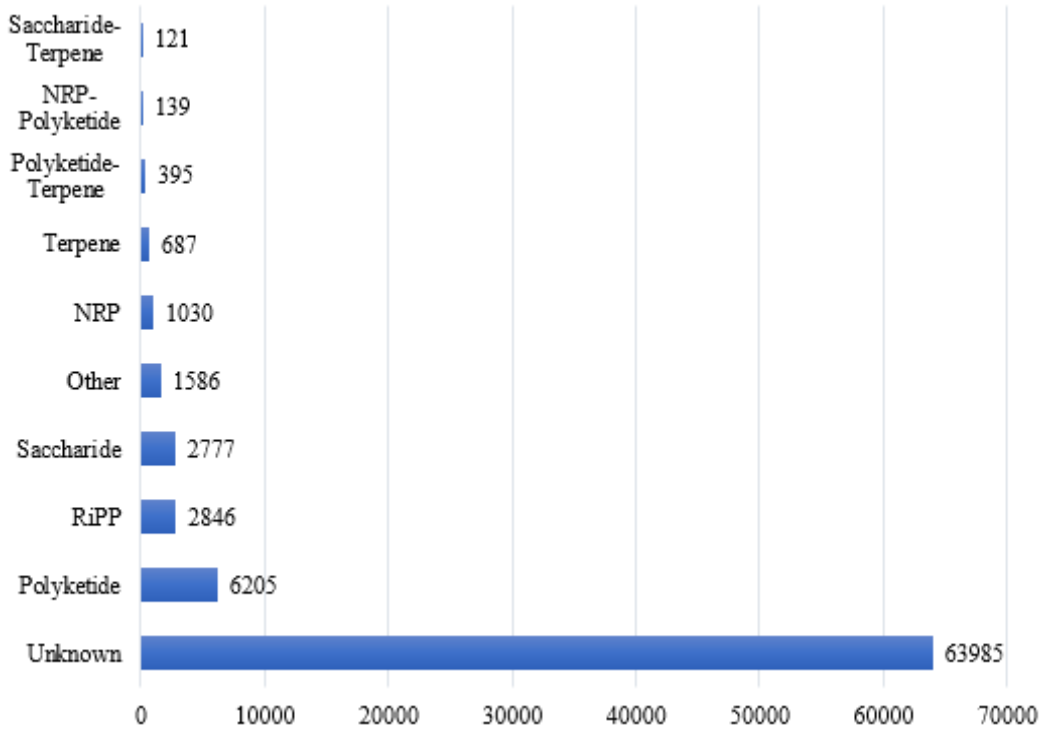

Figure 11. Distribution of the absolute number of the detected BGCs classes by deepBGC in all datasets.

A major advantage of deepBGC is that it could assign product activity to each single detected BGC class with a very high coverage rate. In this study deepBGC assigned product activity to 96% of the hits and the results were 97% of hits have an antibacterial activity, 1% inhibitor, 1% antibacterial-antifungal and less than 1% cytotoxic (Figure 12). About 31% of the detected BGCs were products of microbial species belong to one of the following eight genera, 7% from *Bacteroides*, 5% *Pseudomonas*, 5% *Corynebacterium*, 4% *Gordonia*, 3% *Blautia*, 3% *Escherichia*, 2% *Faecalibacterium* and 2% *Cutibacterium*.

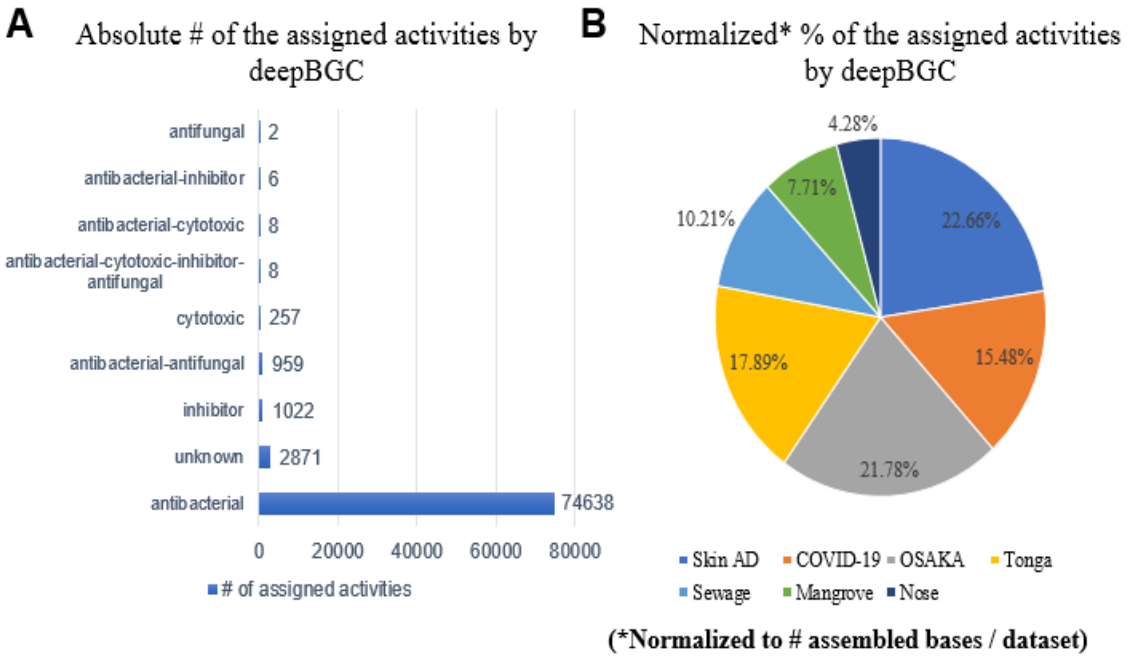

Figure 12. Distribution the assigned product activities by deepBGC in all datasets. (A) Distribution of the absolute number of the assigned products' activities by deepBGC (B) Distribution of the normalized percentages of the assigned products' activities by deepBGC. Percentages were normalized to the number of assembled bases per dataset.

DeepBGC detected 4,795 BGCs from the nine metatranscriptomic marine samples from Osaka Bay see water project. All processed samples gave hits with deepBGC pipeline. DeepBGC assigned BGCs classes to around 17% of the hits (832 hits) as follows; 49% of hits was Polyketide, 13% NRP, 12% RiPP, 8% other, 7% Saccharide, 6% Terpene, 3% Saccharide-Terpene, 2% Polyketide-Terpene and 1% NRP-Polyketide. In this project deepBGC assigned product activity to almost 96% of the hits and the results were 97% of hits have an antibacterial activity, 2% inhibitor and 1% antibacterial-antifungal. DeepBGC specified the microbial species to about 82% of the hits and more than 80% of the detected BGCs were products of microbial species belong to one of the following 3 genera, 63% from *Pseudomonas*, 15% *Synechococcus*, and 4% *Candidatus*.

The results of metagenomic samples from the gut microbiome project of COVID-19 patients were analyzed. In this study, 10 samples were processed, all samples gave hits and deepBGC annotated BGCs classes to around 20% of the hits (6,852 hits) as follows; 30% of hits was Polyketide, 26% Saccharide, 23% RiPP, 11% other, 4% NRP, 3% Terpene, 1% Polyketide-Terpene and 1% NRP-Polyketide. In this project deepBGC assigned product activity to almost 96% of the hits and the results were 98% of hits have an antibacterial activity, 1% inhibitor, 1% antibacterial-antifungal and 1% cytotoxic. DeepBGC specified the microbial species to about 79% of the hits and about 33% of the detected BGCs were products of microbial species

belong to one of the following 4 genera, 16% from *Bacteroides*, 7% *Blautia*, more than 5% *Lachnospiraceae* and 5% *Faecalibacterium*.

A total of 18,683 hits were detected by deepBGC from 10 skin microbiome metagenomic samples of patients with atopic dermatitis. DeepBGC annotated BGCs classes to around 22% of the hits (4,137 hits) as follows; 46% of hits was Polyketide, 14% RiPP, 10% Saccharide, 10% NRP, 8% Other, 5% Terpene, 5% Polyketide-Terpene, 2% Saccharide-Terpene and 1% NRP-Polyketide. Moreover, deepBGC assigned product activity to almost 95% of the hits and 96% of hits have an antibacterial activity, 2% inhibitor and 2% antibacterial-antifungal. DeepBGC specified the microbial species to about 75% of the hits and more than 50% of the detected BGCs were products of microbial species belong to one of the following 5 genera, 16% from *Gordonia*, 14% *Corynebacterium*, 7% *Mycolicibacterium*, 7% *Cutibacterium* and 7% *Micrococcus*.

The five samples from Tonga trench project gave 2,751 hits with deepBGC. The pipeline annotated about 21% of these hits (567 hits) with different BGCs classes as follow; 52% of hits belonged to Polyketide, 12% RiPP, 11% Terpene, 8% Others, 7% NRP, 5% Saccharide, 4% Polyketide-Terpene, 1% Saccharide-Terpene and 1% NRP-Polyketide. On the other hand, 96% of hits annotated by deepBGC with three different activities, the majority about 96% have an antibacterial activity, 3% antibacterial-antifungal and about 1% inhibitor. The pipeline also specified the microbial species to about 45% of the hits and about 20% of the detected BGCs were products of microbial species belong to one of the following 2 genera, 16% *Cutibacterium* and 4% belonged to *Pseudomonas*.

The then metagenomic samples from Mangrove sediment microbiome project gave 1,624 hits with deepBGC. The pipeline annotated about 14% of these hits (230 hits) with different BGCs classes as follow; 46% of hits belonged to Polyketide, 13% Saccharide, 11% NRP, 9% RiPP, 8% Others, 6% Terpene, 5% Polyketide-Terpene and 1% Saccharide-Terpene. On the other hand, 97% of the total detected hits annotated by deepBGC with one of two different activities, the vast majority about 99% have an antibacterial activity and 1% inhibitor. The pipeline also specified the microbial species to about 52% of the hits and about 24% of the detected BGCs were products of microbial species belong to one of the following 3 genera, 10% *Altererythrobacter*, 8% *Erythrobacter* and 6% belonged to *Candidatus Plagibacter*.

The twelve metagenomic samples from a study conducted in Copenhagen; Denmark gave about 1,351 hits with deepBGC. The pipeline annotated about 21% of these hits (283 hits) with different BGCs classes as follow; 45% of hits belonged to Polyketide, 17% RiPP, 12% Other, 9% NRP, 8% Saccharide, 7% Terpene, 1% Polyketide-Terpene and 1% NRP-Polyketide. Moreover, about 95% of the total detected hits annotated by deepBGC with one of three different activities, the vast majority, about 96%, have an antibacterial activity, 2% inhibitor and 1% antibacterial-antifungal. The pipeline also specified the microbial species to about 81% of the hits and the genus *Corynebacterium* was the most dominant as it represents about 78% of the whole microbial community from the processed samples.

The last 9 samples were from the Global Water Sewage Project. DeepBGC pipeline gave 16,682 hits, 17% of it (i.e. 2,813 hits) annotated by deepBGC with different BGCs classes. The majority 44% were Polyketide, 17% RiPP, 15% Saccharide, 12% Other, 5% NRP, 4% Terpene, 2% Polyketide-Terpene, 1% NRP-Polyketide. The pipeline assigned product activity to almost 97% of the total hits and the vast majority, 97%, was antibacterial, 1% was inhibitor and 1% was antibacterial-antifungal. DeepBGC specified the microbial species to about 77% of the hits and about 31% of the detected BGCs were products of microbial species belong to one of the following 6 genera, 9% *Escherichia*, 6% *Acidovorax*, 5% *Neisseria*, 4% *Streptococcus*, 3% *Arcobacter* and the last 3% belonged to *Pseudomonas*.

AMR genes profile of samples in the dataset as detected by CARD's RGI

The second major goal of this study was to detect the antimicrobial resistance genes of the samples from the selected metagenomes, along with their mechanisms of actions and the drug classes which it confers resistance to. Here we used Resistance Gene Identifier (RGI) algorithm from The Comprehensive Antibiotic Resistance Database (CARD) with the following criterions for detection (perfect, strict & loose, partial genes included, 95% identity nudge used) of AMR genes. Loose hits were excluded, here only results of perfect and strict hits were reported to ensure that they are either perfect matches or passed the curated bit-score. In our study, the selected sixty five samples from the different seven selected projects were analyzed by CARD's RGI

and a total number of around 1216 AMR gene families were detected which confer resistance to about 2602 drug classes by 1163 resistance mechanisms. Figures 18 – 20 show the distribution of the detected AMR gene families, drug classes and resistance mechanisms of all samples from the seven selected metagenomes, respectively. Gut microbiome of COVID-19 patients' samples and water sewage samples represent more than 70% of antibiotic resistance abundance, while samples of the rest five selected metagenomes represent less than 30%. To eliminate the effect of the number of bases all percentages were normalized by dividing the total detected number of AMR gene families, drug classes and resistance mechanisms by the total number of used bases per each of the seven selected metagenomes.

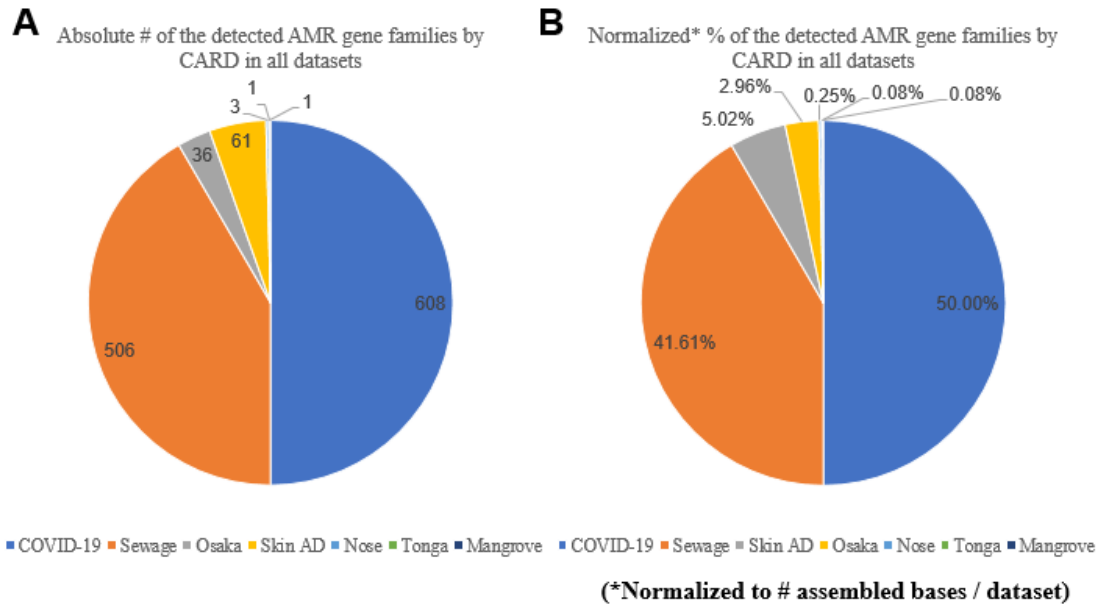

Figure 13. Distribution the detected AMR genes families by CARD in all datasets. (A) Distribution of the absolute number of detected AMR genes families by CARD. (B) Distribution of the normalized percentages of the detected AMR genes families. Percentages were normalized to the number of assembled bases per dataset.

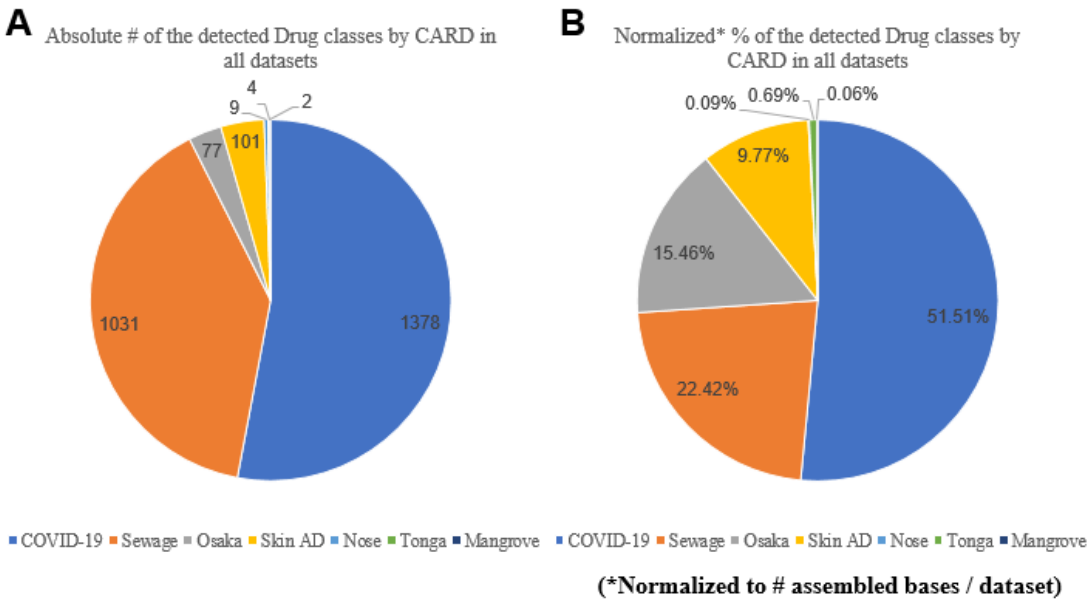

Figure 14. Distribution the detected drug classes by CARD in all datasets. (A) Distribution of the absolute number of detected drug classes by CARD. (B) Distribution of the normalized percentages of the detected drug classes. Percentages were normalized to the number of assembled bases per dataset.

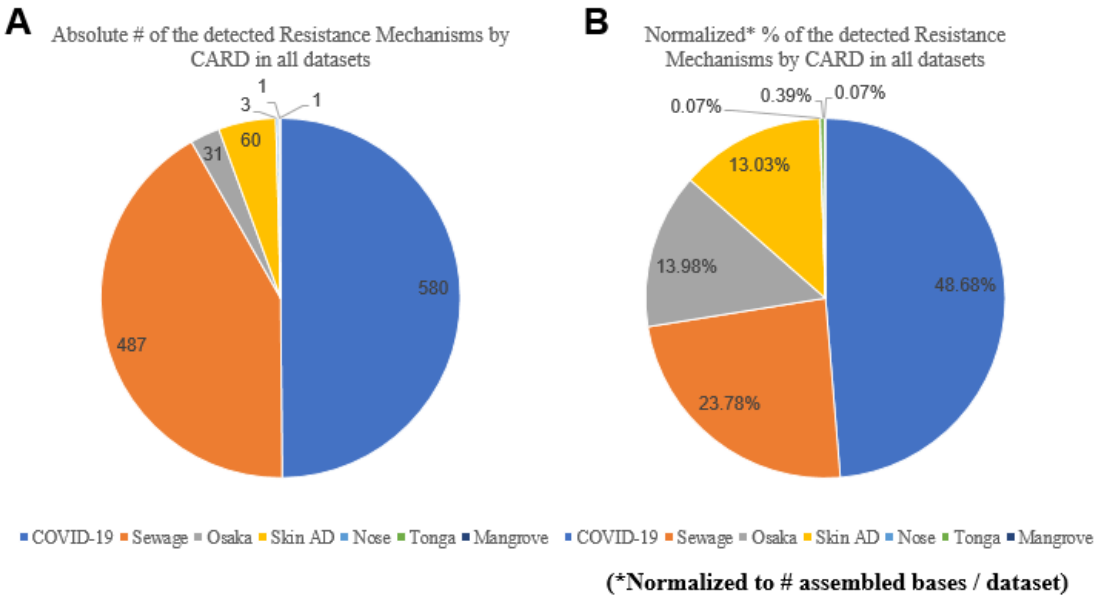

Figure 15. Distribution the detected resistance mechanisms by CARD in all datasets. (A) Distribution of the absolute number of detected resistance mechanisms by CARD. (B) Distribution of the normalized percentages of the detected resistance mechanisms. Percentages were normalized to the number of assembled bases per dataset.

The results of metagenomic samples from the gut microbiome project of COVID-19 patients were analyzed. In this study, 10 samples were processed, 9 samples gave perfect hits (76 hits) while all samples gave strict hits (468 hits). CARD's RGI algorithm detected a total of 608 AMR genes from different 55 families which represent 48.26% of all detected AMR gene families of all processed samples from the seven selected

metagenomes. Such AMR genes confer resistance to a total of 1378 drugs from different 37 classes which represent 51.51% of the overall results (Table 5).

Table 5. Drug classes detected by CARD's RGI in gut microbiome of COVID-19 patients' samples. A total 1378 drug classes from 32 different classes, according to CARD's classification were detected.

#	Drug Classes	Detected #	%
1	tetracycline antibiotic	203	14.73%
2	fluoroquinolone antibiotic	174	12.63%
3	Penam	118	8.56%
4	Cephalosporin	105	7.62%
5	macrolide antibiotic	95	6.89%
6	aminoglycoside antibiotic	73	5.30%
7	rifamycin antibiotic	69	5.01%
8	phenicol antibiotic	59	4.28%
9	glycylcycline	49	3.56%
10	lincosamide antibiotic	44	3.19%
11	Cephamycin	43	3.12%
12	peptide antibiotic	42	3.05%
13	Triclosan	42	3.05%
14	aminocoumarin antibiotic	34	2.47%
15	streptogramin antibiotic	32	2.32%
16	Carbapenem	23	1.67%
17	Penem	23	1.67%
18	diaminopyrimidine antibiotic	20	1.45%
19	nucleoside antibiotic	20	1.45%
20	acridine dye	19	1.38%
21	Monobactam	17	1.23%
22	Fosfomycin	16	1.16%
23	glycopeptide antibiotic	16	1.16%
24	nitrofuran antibiotic	7	0.51%
25	nitroimidazole antibiotic	6	0.44%
26	oxazolidinone antibiotic	6	0.44%
27	pleuromutilin antibiotic	6	0.44%
28	benzalkonium chloride	5	0.36%
29	Rhodamine	5	0.36%
30	elfamycin antibiotic	4	0.29%
31	aminocoumarin antibiotic	2	0.15%
32	sulfonamide antibiotic	1	0.07%
	GRAND TOTAL	**1378**	**100.00%**

These genes confer resistance to the different drug classes by 6 different resistance mechanisms of a total 580 mechanisms, which represent 48.68% of the overall detected resistance mechanisms. The top three detected AMR gene families were, resistance-nodulation-cell division (RND) antibiotic efflux pump (28.62%), major facilitator superfamily (FS) antibiotic efflux pump (17.11%) and tetracycline-resistant ribosomal protection protein (10.36%), which represent 56.09% of the 608 detected AMR gene families. Whereas, the top first five drug classes which represent 50.44% of the detected 1378 drug classes were as follows, tetracycline antibiotic (14.73%), fluoroquinolone antibiotic (12.63%), penam (8.56%), cephalosporin (7.62%) and macrolide antibiotic (6.89%). On the other hand, the six detected resistance

mechanisms by which genes introduce drug resistance, were as follows; antibiotic efflux (47.93%), antibiotic target alteration (20%), antibiotic inactivation (13.28%), antibiotic target protection (12.24%), antibiotic target replacement (3.97%) and reduced permeability to antibiotic (2.59%).

The nine metagenomic samples from the Global Water Sewage Project come in the second place after gut microbiome of COVID-19 patients' samples in terms of antibiotic resistance abundance. All samples gave hits with RGI where; 96 of hits were perfect matches and 370 of the hits were strict. A total 506 AMR genes belong to 37 different families was detected, which represent 23.37% of all detected AMR gene families. These genes confer resistance to 1031 drugs belong to 29 different classes which collectively represent 22.42% of the whole detected drug classes (Table 6).

Table 6. Drug classes detected by CARD's RGI in water sewage samples. A total 1031 drug classes from 29 different classes, according to CARD's classification were detected.

#	Drug Classes	Detected #	%
1	tetracycline antibiotic	155	15.03%
2	Penam	123	11.93%
3	fluoroquinolone antibiotic	103	9.99%
4	macrolide antibiotic	90	8.73%
5	Cephalosporin	87	8.44%
6	aminoglycoside antibiotic	77	7.47%
7	phenicol antibiotic	49	4.75%
8	Cephamycin	37	3.59%
9	rifamycin antibiotic	35	3.39%
10	Glycylcycline	30	2.91%
11	Triclosan	29	2.81%
12	aminocoumarin antibiotic	25	2.42%
13	peptide antibiotic	22	2.13%
14	sulfonamide antibiotic	20	1.94%
15	acridine dye	18	1.75%
16	nucleoside antibiotic	17	1.65%
17	diaminopyrimidine antibiotic	15	1.45%
18	lincosamide antibiotic	15	1.45%
19	streptogramin antibiotic	15	1.45%
20	Monobactam	14	1.36%
21	Penem	14	1.36%
22	oxazolidinone antibiotic	9	0.87%
23	pleuromutilin antibiotic	9	0.87%
24	Carbapenem	7	0.68%
25	nitroimidazole antibiotic	5	0.48%
26	Fosfomycin	4	0.39%
27	Carbapenem	3	0.29%
28	benzalkonium chloride	2	0.19%
29	Rhodamine	2	0.19%
	GRAND TOTAL	**1031**	**100.00%**

These genes confer resistance to the different drug classes by 6 different resistance mechanisms of a total 487 mechanisms, which represent 23.78% of the overall detected resistance mechanisms. Three major AMR gene families which represent 55.73% of 506 detected AMR genes were as follows; major facilitator superfamily (MFS) antibiotic efflux pump (24.90%), resistance-nodulation-cell division (RND) antibiotic efflux pump (20.75%) and tetracycline-resistant ribosomal protection protein (10.08%). Out of the 29 different drug classes, there were 6 major classes which represent 61.59% from the total number of 1031 detected drugs, and they were as follows; tetracycline antibiotic (15.03%), penam (11.93%), fluoroquinolone antibiotic (9.99%), macrolide antibiotic (8.73%), cephalosporin (8.44%) and aminoglycoside antibiotic (7.47%). As in gut microbiome of COVID-19 patients' samples, there were six different resistance mechanisms detected but with different percentages as follows; antibiotic efflux (45.38%), antibiotic inactivation (27.72%), antibiotic target protection (13.14%), antibiotic target replacement (6.78%), antibiotic target alteration (6.37%) and reduced permeability to antibiotic (0.62%).

The results of the rest five selected metagenomes represent less than 30% (normalized value) in terms of AMR gene family, drug classes and resistance mechanisms. No perfect hits were detected from the nine metatranscriptomic marine samples from Osaka Bay see water project while there were 26 strict hits from 5 samples detected. A total of 36 AMR genes from different 5 families which represent 15.35% of all detected AMR gene families of all processed samples from the seven selected metagenomes. The detected AMR genes confer resistance to a total of 77 drugs from different 11 classes which represent 15.46% of the overall results (Table 7).

Table 7. Drug classes detected by CARD's RGI in Osaka bay samples. A total 77 drug classes from 11 different classes, according to CARD's classification were detected.

#	Drug Classes	Detected #	%
1	aminoglycoside antibiotic	10	12.99%
2	fluoroquinolone antibiotic	15	19.48%
3	tetracycline antibiotic	15	19.48%
4	triclosan	6	7.79%
5	cephalosporin	5	6.49%
6	glycylcycline	5	6.49%
7	penam	5	6.49%
8	acridine dye	5	6.49%
9	rifamycin antibiotic	5	6.49%
10	phenicol antibiotic	5	6.49%
11	macrolide antibiotic	1	1.30%
	GRAND TOTAL	**77**	**100.00%**

The detected AMR genes confer resistance to the different drug classes by 3 different resistance mechanisms of a total 31 mechanisms, which represent 13.98% of the overall detected resistance mechanisms. The five detected AMR gene families were, resistance-nodulation-cell division (RND) antibiotic efflux pump (41.67%), major facilitator superfamily (MFS) antibiotic efflux pump (16.67%), APH(3") (13.89%), APH(6) (13.89%) and ATP-binding cassette (ABC) antibiotic efflux pump (13.89%), which all represents 14.52% of the 36 detected AMR gene families. However, the first major three drug classes which represent 51.95% of the 77 detected drug classes were as follows, tetracycline antibiotic (19.48%), fluoroquinolone antibiotic (19.48%) and aminoglycoside antibiotic (12.99%). The three detected resistance mechanisms by which genes introduce drug resistance, were as follows; antibiotic efflux (51.61%), antibiotic inactivation (32.26%) and antibiotic target alteration (16.13%).

On the other hand, from the ten samples of Skin AD metagenomes, only 4 samples gave 7 perfect hits whereas all samples gave strict hits and the total detected number was 53 hits. The processed samples yielded a total number of 61 AMR genes from different 25 families which represent 12.53% of all detected AMR gene families compared to the rest of samples from the seven selected projects. These AMR genes confer resistance to a total of 101 drugs from different 23 classes which represent 9.77% of the overall results (Table 8).

Table 8. Drug classes detected by CARD's RGI in skin AD patients' samples. A total 101 drug classes from 23 different classes, according to CARD's classification were detected.

#	Drug Classes	Detected #	%
1	aminoglycoside antibiotic	15	14.85%
2	macrolide antibiotic	11	10.89%
3	lincosamide antibiotic	11	10.89%
4	streptogramin antibiotic	11	10.89%
5	fluoroquinolone antibiotic	9	8.91%
6	phenicol antibiotic	7	6.93%
7	Penam	6	5.94%
8	tetracycline antibiotic	5	4.95%
9	peptide antibiotic	4	3.96%
10	oxazolidinone antibiotic	3	2.97%
11	pleuromutilin antibiotic	3	2.97%
12	fusidic acid	2	1.98%
13	lincosamide antibiotic	2	1.98%
14	diaminopyrimidine antibiotic	2	1.98%
15	acridine dye	2	1.98%
16	aminocoumarin antibiotic	1	0.99%
17	elfamycin antibiotic	1	0.99%
18	Cephalosporin	1	0.99%
19	Fosfomycin	1	0.99%
20	glycopeptide antibiotic	1	0.99%
21	Mupirocin	1	0.99%
22	rifamycin antibiotic	1	0.99%
23	sulfonamide antibiotic	1	0.99%
	GRAND TOTAL	**101**	**100.00%**

The detected AMR genes confer resistance to the different drug classes by 5 different resistance mechanisms of a total 60 mechanisms, which represent 13.03% of the overall detected resistance mechanisms. The most abundant detected AMR gene families which represent 52.46% of all detected AMR genes were as follows, major facilitator superfamily (MFS) antibiotic efflux pump (18.03%), Erm 23S ribosomal RNA methyltransferase (13.11%), blaZ beta-lactamase (8.20%), APH(3") (6.56%) and APH(6) (6.56%). However, the first major five drug classes which represent 56.44% of the 101 detected drug classes were as follows, aminoglycoside antibiotic (14.85%), macrolide antibiotic (10.89%), lincosamide antibiotic (10.89%), streptogramin antibiotic (10.89%) and fluoroquinolone antibiotic (8.91%). While the five detected resistance mechanisms by which genes introduce drug resistance, were as follows; antibiotic inactivation (35%), antibiotic target alteration (28.33%), antibiotic efflux (23.33%), antibiotic target protection (10%) and antibiotic target replacement (3.33%).

The next samples from the last three selected metagenomes represent the smallest fraction of all results. Their combined results gave less than 1% compared to the rest of results. No perfect hits were detected and only few strict hits were reported as follows, 3 hits, 1 hit and 1 hit from Nose, Tonga and Mangrove projects, respectively. These results reflect a few number of detected AMR gene families (i.e. 3, 1 and 1 for each projects on the same stated order, Nose, Tonga and Mangrove), which represent only 0.49% from the whole results. The detected AMR gene families were as follows, 3 Erm 23S ribosomal RNA methyltransferase, 1 TEM beta-lactamase and 1 resistance-nodulation-cell division (RND) antibiotic efflux pump for samples of Nose, Tonga and Mangrove projects, respectively. Regarding the detected drug classes, we reported a total of 15 drug classes from the three projects, Nose samples came in the first place by 9 classes, then Tonga with 4 classes and last place was for Mangrove samples with only 2 classes. These 15 drug classes represent 0.84% from the overall results (Tables 9 – 11). The nine drug classes of Nose project were from three different classes, macrolide antibiotic, lincosamide antibiotic and streptogramin antibiotic which share the same percentage 33%. While the four detected drug classes of Tong project were from four different classes as follows; monobactam, cephalosporin, penam and penem. We also reported only 2 drug classes from samples of Mangrove project as follows; fluoroquinolone antibiotic and tetracycline antibiotic. Five resistance mechanisms, represent 0.52%, were also reported from the three projects, 3 from Nose, 1 from Tonga

and 1 from Mangrove. These five mechanisms belong to three different types as follows; antibiotic target alteration, antibiotic inactivation and antibiotic efflux for Nose, Tonga and Mangrove metagenomes, respectively.

Table 9. Drug classes detected by CARD's RGI in nose samples. A total 9 drug classes from 3 different classes, according to CARD's classification were detected.

#	Drug Classes	Detected #	%
1	macrolide antibiotic	3	33%
2	lincosamide antibiotic	3	33%
3	streptogramin antibiotic	3	33%
	GRAND TOTAL	**9**	**100.00%**

Table 10. Drug classes detected by CARD's RGI in Tonga trench samples. A total 4 drug classes from 4 different classes, according to CARD's classification were detected.

#	Drug Classes	Detected #	%
1	Monobactam	1	25%
2	Cephalosporin	1	25%
3	Penam	1	25%
4	Penem	1	25%
	GRAND TOTAL	**4**	**100.00%**

Table 11. Drug classes detected by CARD's RGI in Mangrove samples. A total 2 drug classes from 2 different classes, according to CARD's classification were detected.

#	Drug Classes	Detected #	%
1	fluoroquinolone antibiotic	1	50%
2	tetracycline antibiotic	1	50%
	GRAND TOTAL	**2**	**100.00%**

Chapter 4: Discussion

According to WHO, antimicrobial resistance is considered one of the most complex global health challenges and should be a political priority. Dr Chan, WHO Former Director-General, said "The World Bank has warned that antimicrobial resistance could cause as much damage to the economy as the 2008 financial crisis." Moreover, with the limited choices of replacement products, WHO experts expect that the world is heading toward a post antibiotic era and the common infectious diseases will be mortal once again. This put a great pressure on the scientific community all over the world to discover new classes of antibiotics with new mechanisms of actions against the rising number of antimicrobial resistant bacterial strains.

Over the third of small molecule medicines approved by the FDA were microbial natural products encoded by neighboring genes called Biosynthetic Gene Clusters (BGCs) (Newman & Cragg, 2020 & Martin, 1992). To date, the gold standard way for discovering bioactive natural products is the culture-dependent techniques which considered a major challenge due to the fact that only small fraction of bacterial species could be cultured under the current laboratory conditions (Stewart, 2012). The advancement of sequencing technologies and omics approaches unleash the power of natural products discovery through exploring the uncultivated microbial species which represent the biggest fraction of microbial community. Here we tried to positively contribute in solving the antimicrobial resistance problem by two different ways. Firstly, we tried to support researchers who interested in natural products discovery through catalog the bacterial BGCs in the selected metagenomes by conducting a comparative analysis to determine the different BGCs' classes present in each of the selected samples along with highlighting the major bacterial species contributors. Secondly, we did a thorough analysis to detect the antimicrobial resistance genes with their resistance mechanisms along with the drug classes they confer resistance to, in order to shed the light on this crises with deeper insights.

Different environments have different microbial taxa profile

Each environments comprise huge microbial communities live in complex interactions that greatly impact our life. Therefore, such comparative studies aiming to precisely profile the microbial communities' compositions and their corresponding contributions in terms of production of secondary metabolites, are of fundamental

interest. Our results show that the different environmental conditions are the major determinants of the microbial composition. By screening 65 samples from seven different metagenomic and metatranscriptomic projects we discovered reads belong to a wide range of different microbial taxa, at the genus level, and some of them was found to be unique and characteristic to their corresponding environments. For example, samples from Osaka bay project were characterized by the presence of three major genera, in terms of their BGC contribution, the first place goes to *Pseudomonas* which contribute alone by more than 60% of the detected BGCs, *Synechococcus* comes in the second place by more than 15% while the third place goes to *Candidatus Plagibacter* by about 4% of the detected BGCs. Moreover, *Synechococcus* was not recognized in any other samples from the other 6 projects therefor it was unique and characteristic to this environment at the time of sampling. Mangrove samples, on the other hand, were characterized by the presence of two unique genera which not reported elsewhere in our study, *Altererythrobacter* and *Erythrobacter* which contributed by about 10% and 8% of the detected BGCs respectively. *Candidatus*, which appeared to be characteristic also to Osaka bay environment, comes here in the third place contributing by about 6% of the detected BGCs (Figure 16).

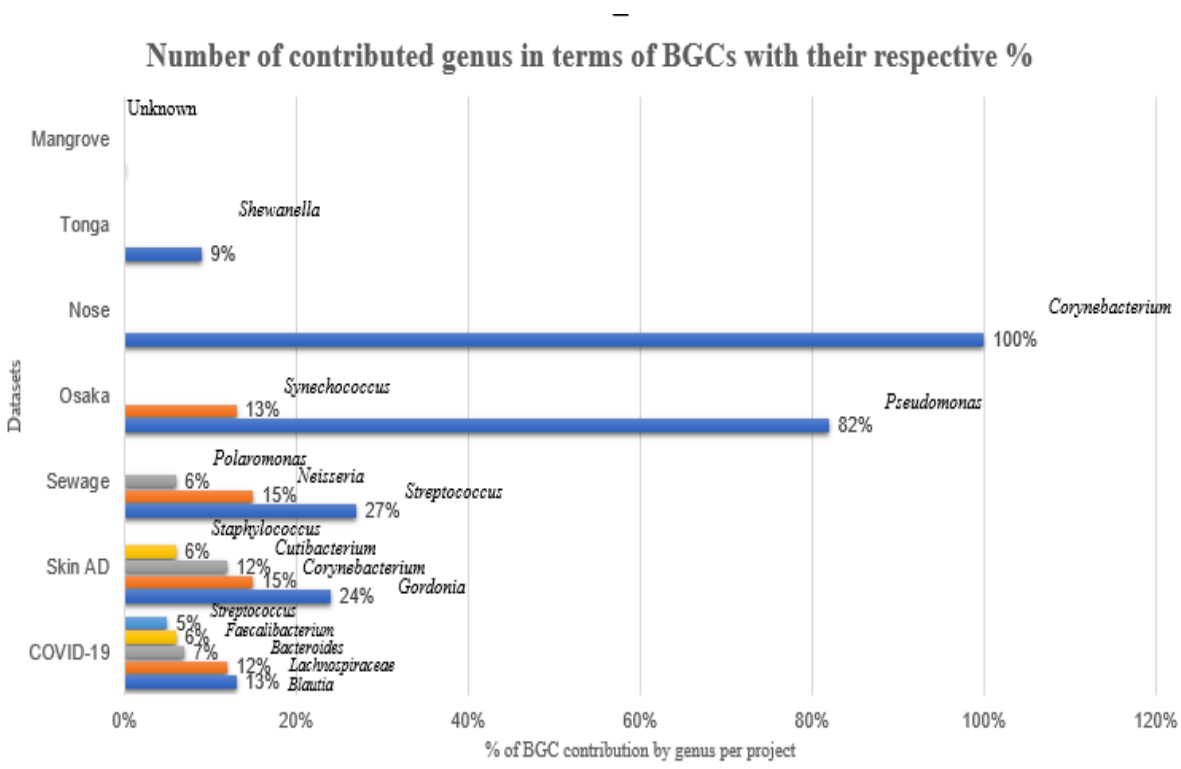

Figure 16. Distribution of the number of contributed genus to BGCs with their respective contribution percentages as assigned by antiSMASH per dataset.

Table 12 has a complete comparison between antiSMASH and deepBGC results in terms of the detected BGC classes with their corresponding percentages and the most abundant genera with their contributions' percentages. It was obvious that each environment had a signature microbial taxa profile, at the time of sampling, in terms of the relative abundance and sometimes there is a unique genera specific to each particular environments. In Tonga trench samples two genera were contributed the most to the detected BGCs, *Cutibacterium* (~16%) and *Pseudomonas* (~4%) while *Shewanella* was unique to this environment, although, it present in a relatively low abundance it survived and this case and other similar cases will be explained with some details in the next sections.

Table 12. Comparison between antiSMASH and deepBGC in terms of the detected BGCs' classes with their corresponding percentage and the most abundant genera with their percentage of contribution.

Projects	**antiSMASH**				**deepBGC**			
	BGC classes	**%**	**Genus**	**%**	**BGC classes**	**%**	**Genus**	**%**
Osaka Bay	NRPS	44%	*Pseudomonas*	82%	Polyketide	49%	*Pseudomonas*	63%
	Bacteriocin	20%	*Synechococcus*	13%	NRP	13%	*Synechococcus*	15%
	Terpene	9%			RiPP	12%	*Candidatus Plagibacter*	4%
	Arylpolyene	7%			Other	8%		
	NRPS-like	7%			Saccharide	7%		
	Siderophore	5%			Terpene	6%		
	hserlactone	5%			Saccharide-Terpene	3%		
	betalactone	3%			Polyketide-Terpene	2%		
	NAGGN	1%			NRP-Polyketide	1%		
COVID-19	sactipeptide	28%	*Blautia*	13%	Polyketide	30%	*Bacteroides*	16%
	NRPS	19%	*Lachnospiraceae*	12%	Saccharide	26%	*Blautia*	7%
	bacteriocin	16%	*Bacteroides*	7%	RiPP	23%	*Lachnospiraceae*	5%
	arylpolyene	12%	*Faecalibacterium*	6%	Other	11%	*Faecalibacterium*	5%
	lanthipeptide	7%	*Streptococcus*	5%	NRP	4%		
	NRPS-like	5%			Terpene	3%		
	Terpene	2%			Polyketide-Terpene	1%		
	T3PKS	2%			NRP-Polyketide	1%		
	lassopeptide	2%			Saccharide-Terpene	0.13 %		
	Betalactone	2%						
	Resorcinol	2%						
	Siderophore	2%						
	Thiopeptide	0.4%						
	Nucleoside	0.4%						
	butyrolactone	0.4%						
	Ladderane	0.4%						
Skin (AD)	NRPS	27%	*Gordonia*	24%	Polyketide	46%	*Gordonia*	16%
	NRPS-like	17%	*Corynebacterium*	15%	RiPP	14%	*Corynebacterium*	14%
	Siderophore	10%	*Cutibacterium*	12%	Saccharide	10%	*Mycolicibacterium*	7%
	Terpene	10%	*Staphylococcus*	6%	NRP	10%	*Cutibacterium*	7%
	Bacteriocin	9%			Other	8%	*Micrococcus*	7%
	T1PKS	8%			Terpene	5%		
	Ectoine	4%			Polyketide-Terpene	5%		
	T3PKS	3%			Saccharide-Terpene	2%		
	Hserlactone	3%			NRP-Polyketide	1%		

	Thiopeptide	2%						
	Betalactone	1%						
	lanthipeptide	1%						
	Arylpolyene	1%						
	CDPS	1%						
	LAP	1%						
	hglE-KS	1%						
	Ladderane	1%						
	butyrolactone	0.4%						
	lassopeptide	0.4%						
Tonga	Arylpolyene	17%	*Shewanella* Unique genus and produce unique product (phosphonate)	NA	Polyketide	52%	*Cutibacterium*	16%
	NRPS-like	17%			RiPP	12%	*Pseudomonas*	4%
	hglE-KS	17%			Terpene	11%		
	Bacteriocin	13%			Other	8%		
	phosphonate	9%			NRP	7%		
	Terpene	9%			Saccharide	5%		
	Hserlactone	9%			Polyketide-Terpene	4%		
	T1PKS	4%			Saccharide-Terpene	1%		
	NRPS	4%			NRP-Polyketide	1%		
Mangrove	Bacteriocin	33%	-	-	Polyketide	46%	*Altererythrobacter*	10%
	Arylpolyene	17%			Saccharide	13%	*Erythrobacter*	8%
	NRPS-like	17%			NRP	11%	*Candidatus Plagibacter*	6%
	Terpene	17%			RiPP	9%		
	lassopeptide	8%			Other	8%		
	NRPS	8%			Terpene	6%		
					Polyketide-Terpene	5%		
					Saccharide-Terpene	1%		
					NRP-Polyketide	0.4%		
Skin (Nose)	NRPS	39%	*Corynebacterium*	100 %	Polyketide	45%	*Corynebacterium*	78%
	Terpene	22%			RiPP	17%		
	Siderophore	22%			Other	12%		
	NRPS-like	9%			NRP	9%		
	T1PKS	9%			Saccharide	8%		
					Terpene	7%		
					Polyketide-Terpene	1%		
					NRP-Polyketide	1%		
Sewage	Bacteriocin	30%	*Streptococcus*	27%	Polyketide	44%	*Escherichia*	9%
	Terpene	28%	*Neisseria*	15%	RiPP	17%	*Acidovorax*	6%
	hserlactone	10%	*Polaromonas*	6%	Saccharide	15%	*Neisseria*	5%
	NRPS-like	8%			Other	12%	*Streptococcus*	4%
	arylpolyene	7%			NRP	5%	*Arcobacter*	3%
	sactipeptide	5%			Terpene	4%	*Pseudomonas*	3%
	resorcinol	3%			Polyketide-Terpene	2%		
	T3PKS	3%			NRP-Polyketide	1%		
	Ectoine	2%			Saccharide-Terpene	0.3%		
	butyrolactone	2%						
	phenazine	1%						
	RaS-RiPP	1%						

It was expected that similar environments in terms of their nature, most probably would have similar microbial composition. This was reported in our study as follows, samples from two different skin environments were addressed; samples of the Human skin metagenome from epithelium of external nose project were characterized and dominated by *Corynebacterium* while the samples from metagenomic skin of patients with AD were characterized by the following genera, *Corynebacterium* and *Cutibacterium*. Here *Corynebacterium* was the first contributors in both skin environments by more than 75% and about 14% of the detected BGCs in nose & skin AD samples respectively. *Cutibacterium* comes in the second place in terms of BGCs contribution by about 7% in skin AD patients samples and this might be due to an arm race between *Corynebacterium* and *Cutibacterium* and both genera were trying to create their own niche at the time of sampling. Three water in nature environments, Osaka, Tonga and Sewage, also were characterized by the presence of *Pseudomonas* genera with a high relative abundance in Osaka bay and very low abundance in both Tonga and Sewage samples. Moreover, samples from both COVID-19 patients and sewage represent gut microbiome community and this could explain the presence of *Streptococcus* genera in both samples. The Barplot in Figure 4 in results section, showing the relative abundance of all the detected microbial taxa, at the genus level, of the 65 processed samples per each of the seven selected projects.

Our analysis also shows that there might be common bacterial strains between irrelevant environments, such as the presence of *Cutibacterium* in samples from both skin AD patients and Tonga trench. Moreover, we reported the presence of *Candidatus* in both samples from Osaka bay and Mangrove with relatively low abundance in both. To show how the different samples will be clustered based on the relative abundance of taxa we constructed both a PCA and t-SNE graphs (Figures 5 & 6 in results section). Projects like Osaka bay and water sewage which dominated by *Pseudomonas* and *Streptococcus* respectively with a high abundance, appeared completely separated from the rest five projects, while there were many connections between the other projects, which mainly due to the presence of common genera, such as *Corynebacterium* and *Cutibacterium* which explained the presence of some samples, appeared on both figures as colored dots, from the Human skin metagenome from epithelium of external nose and Tonga trench project around the samples of the skin project of AD patients (Table 12).

It was also obvious that samples of some projects were dominated by few unique genera such as the samples from Osaka bay project which were dominated mainly by *Pseudomonas* with very large percentage followed by *Synechococcus* with relatively high percentage. Moreover, samples from the project of metagenomic skin of patients with AD were dominated mainly by *Cutibacterium* which is characteristic to skin environment, while *Corynebacterium* stands alone in the samples of the Human skin metagenome from epithelium of external nose project. Species like *Streptococcus agalactiae* were unique and characteristic to the samples of the water sewage projects. Although *Streptococcus agalactiae* present in a very low abundance it survived at the time of sampling. To understand the reason behind the presence of different microbial profiles in each environments, where some genera stands alone in some samples while there are samples with many different species live together and how could some species survive with a very low abundance, we analyzed the microbial biosynthetic potential of each environment on the next section.

The biosynthetic potential of the selected microbial metagenomes

The importance of secondary metabolites came from their potential applications, as assigned by deepBGC the vast majority, about 95%, of the detected BGCs have an antibacterial activity. According to antiSMASH results, 45 out of 65 samples from the seven selected projects give hits of a total 776 BGCs regions of different 26 classes. We found that the number of BGCs and their classes directly proportional with the degree of microbial diversity (Figure 17).

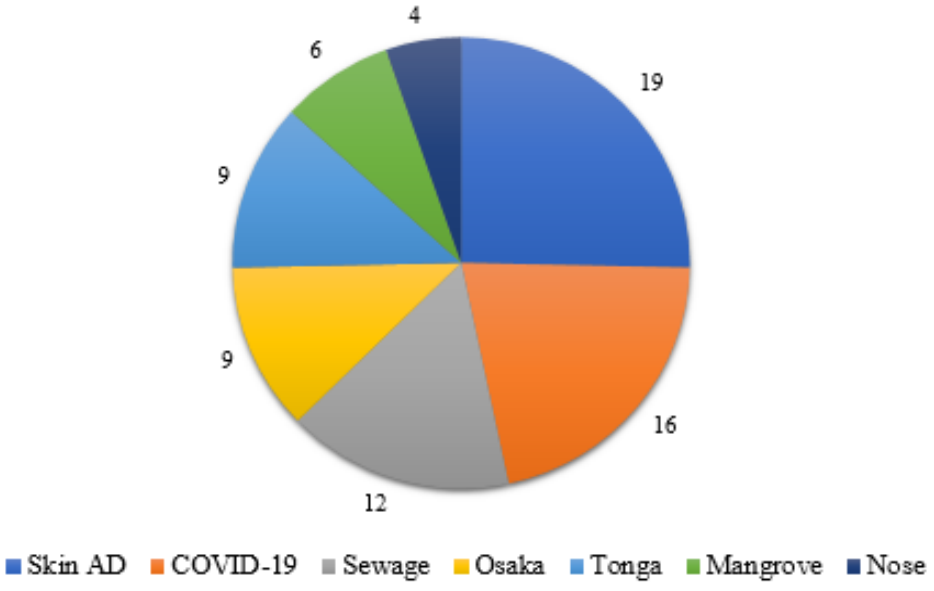

Figure 17. Distribution absolute number the different BGCs classes detected by antiSMASH per dataset.

Environments with a high degree of microbial diversity such as skin AD, gut microbiome of COVID-19 patients and sewage were very rich in terms of total number of detected BGCs and also in the number of BGCs' classes. Ten samples from skin AD come in the first place with 272 detected BGCs belong to 19 different classes with 2 unique classes (i.e. CDPS & LAP) which not reported elsewhere in our study. About 57% of the detected BGCs belongs to four genera, *Gordonia*, *Corynebacterium*, *Cutibacterium* and *Staphylococcus*. There might be an arm race between these genera and each one trying to use as many weapons (i.e. SMs) as they can to create their own niche. Figure 18 showing BGC hits detected by antiSMASH boxplot for each dataset, in relation to its assigned genus.

Figure 18. BGC hits detected by antiSMASH boxplot for each dataset, in relation to its assigned genus.

The majority of the detected BGCs has antimicrobial activity such as NRPS, represents 27% of BGCs, which was reported to have antibacterial activity and one major example of this group is β-lactams and some has also an antitumor activity (Felnagle et al., 2008). Terpene, bacteriocin, T1PKS, T3PKS, thiopeptide, betalactone, lanthipeptide, LAP and lassopeptide are also examples of such classes with an antimicrobial activity. On the other hand there should be a dialog and some sort of coordination between the community members through quorum sensing, this could be explained by the presence of homoserine lactone cluster (hserlactone) which known to has a rule in quorum sensing (Churchill et al., 2011). Butyrolactone also was detected which considered a type of signaling molecules that manage group of genes involved in the bacterial specialized metabolism and morphological differentiation (Horinouchi et al., 2001). Table 13 has more example of the potential use of the secondary metabolites detected in our analysis.

Table 13. Potential application of some detected secondary metabolites

BGC class	Potential application	Reference
NRPS	The majority has antibacterial activity (e.g. β-lactams) and antitumor effect (e.g. bleomycin)	Felnagle et al., 2008
Saccharides	Some have antibacterial activity	Weitnauer et al., 2001
Terpene	Subgroup of terpenes have antibacterial activity	Brahmkshatriya & Brahmkshatriya, 2013
Polyketides (T1PKS)	Subgroup of T1PKS are involved in antibiotic synthesis (e.g. erythromycin)	Yu et al., 2012
Polyketides (T3PKS)	Antibacterial and antitumor activity	Lim et al., 2016
Phosphonate	Have antibacterial activity (e.g. fosfomycin)	Metcalf & van der Donk, 2009
Ectoine	Have a potential use in prevention of Alzheimer's	Jorge et al., 2016
Ras-RiPP	Can produce peptides involved the control of a quorum sensing (QS) system	Ye et al., 2020
Phenazine	Has a role as cell signals that regulate patterns of gene expression	Pierson & Pierson, 2010
Bacteriocin	Peptidic toxins inhibit the growth of similar or closely related bacterial strains	Cotter et al., 2013
Arylpolyene	Antioxidants which protect the bacteria from reactive oxygen species.	Carter, J.; et al., 2016
Siderophore	Responsible mainly for iron transportation across cell membranes	Cornelis P & Andrews SC, 2010
Hserlactone	Quorum sensing	Churchill et al., 2011
NAGGN	Contribute to bacterial cell survival	Matthias Kurz et al., 2010
RiPP	Has more than 20 sub-classes with many applications (i.e. Antibiotics, food preservative, animal feed additives and in cell biology anantin is used as an atrial natriuretic peptide receptor inhibitor)	Arnison PG et al., 2013 Wyss DF et al., 1993
Betalactone	β-lactones appear in different NP classes, such as PKs, nonribosomal peptides and terpenoids. It has inhibition activities for ligases, transferases, oxidoreductases and hydrolases.	Robinson et al., 2018 Lehmann et al., 2018
Sactipeptide	A member of bacteriocin class I which has antimicrobial activity.	Arnison PG et al., 2013
Lanthipeptide	A member of bacteriocin class I which has antimicrobial activity.	Arnison PG et al., 2013
Lassopeptide	A member of bacteriocin class I which has antimicrobial activity.	Arnison PG et al., 2013

Resorcinol	Has a structural roles in membrane formation and associated with wide biological activities such as antibacterial, cytotoxic, dermatotoxic, antioxidant and genotoxic	H. Kikuchi et al., 2017
Thiopeptide	Has antimicrobial activity against several drug-resistance pathogens	R Liao et al., 2009
Nucleoside	Could inhibit bacterial RNA polymerase and has antibacterial activity against drug-resistance bacteria	SI Maffioli et al., 2017
Butyrolactone	Type of signaling molecule that manages group of genes involved in the bacterial specialized metabolism and morphological differentiation.	Horinouchi et al., 2001
Ladderane	Potential biofuel	Javidpour, P et al., 2016
LAP	Has antibacterial activity	DY Travin et al., 2019
hglE-KS	Type of Polyketide synthases (PKS) which has many pharmaceutical activities such as antibacterial, antifungal & antitumor.	Jenke-Kodama et al., 2005

Another example of such complex and diverse environment is COVID-19 patients' samples. Five genera, *Blautia*, *Lachnospiraceae*, *Bacteroides*, *Faecalibacterium* and *Streptococcus* contribute by about 43% from a total number of 243 detected BGCs belong to 16 different classes. Nucleoside which could inhibit bacterial RNA polymerase and has antibacterial activity against drug-resistance bacteria, was unique to this environment (SI Maffioli et al., 2017). Gut microbiome of COVID-19 patients' samples also were characterized by the presence of a wide array of bacteriocin, peptidic toxins which have the ability to inhibit the growth of similar or closely related bacterial strains (Cotter et al., 2013) in addition to the presence of sactipeptide, lanthipeptide and lassopeptide which considered members of bacteriocin class I which have antimicrobial activity (Arnison PG et al., 2013). Moreover, thiopeptide was also detected which has an antimicrobial activity against several drug-resistance pathogens (R Liao et al., 2009). Figure 19 is a heatmap for each dataset with each of the BGC hits from antiSMASH in relation to its assigned genus.

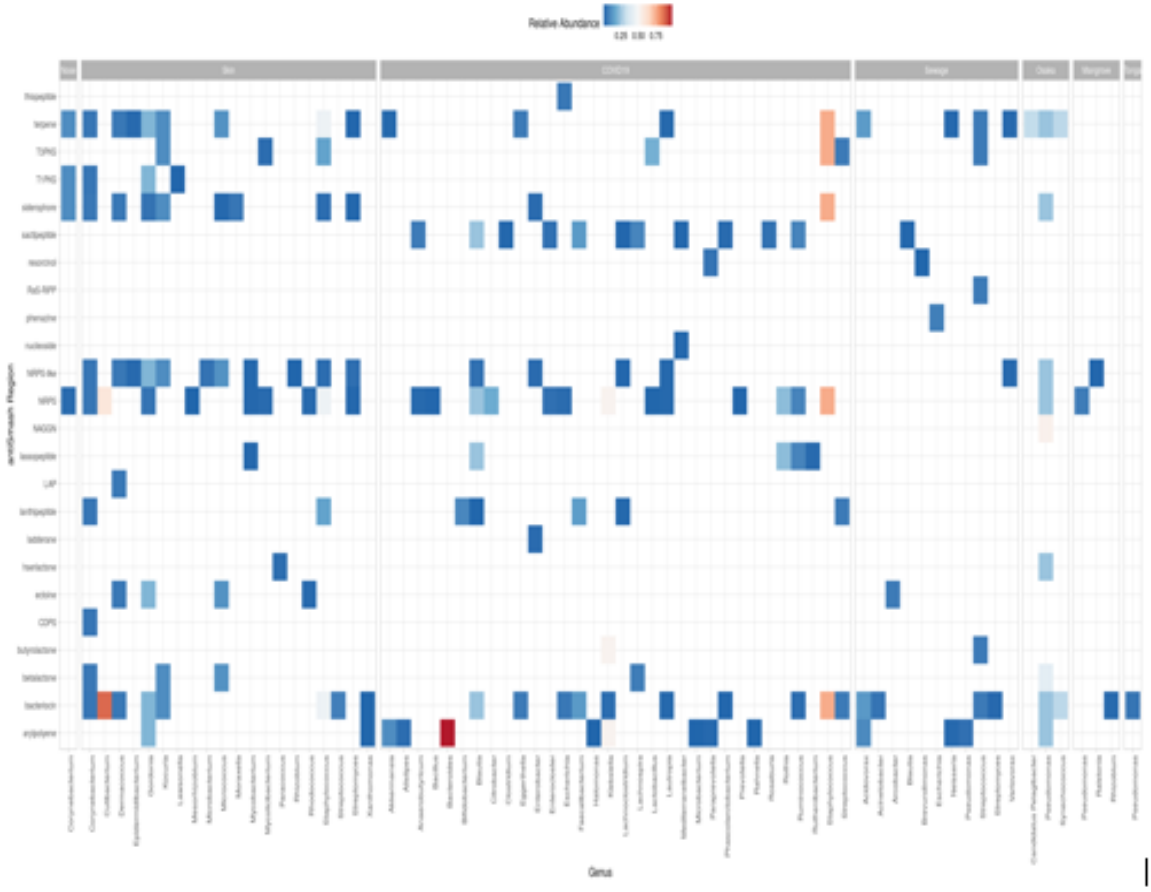

Figure 19. Heatmap for each dataset with each of the BGC hits from antiSMASH in relation to its assigned genus.

Sewage samples comes in the third place in terms of the number of detected BGCs with a total of 102 BGCs belonging to 12 different classes. Here about 50% of the detected BGCs comes from 3 different genera, *Streptococcus, Neisseria* and *Polaromonas.* Different BGCs classes with antimicrobial activity were detected such as bacteriocin, terpene, NRPS-like, sactipeptide and T3PKS. This environment was characterized also by the presence of many signaling and regulating classes such as hserlactone, butyrolactone, phenazine and RaS-RiPP. The last two classes were unique to this environments where phenazine, has a role as cell signals that regulate patterns of gene expression (Pierson & Pierson, 2010) while RaS-RiPP, a product of *Streptococcus Agalactiae,* can produce peptides involved in the control of a quorum sensing (QS) system (Ye et al., 2020). On the other hand, phenazine could be a good example to prove that some species might produce new metabolites under different environmental conditions, in this study we noticed that *Escherichia coli* from sewage water samples only produce phenazine in such environment and we didn't recognize this elsewhere from any other projects.

Many other BGCs' classes were detected in other samples such as saccharides which also have antibacterial activity especially the subset which has a cellular diffusible ability (Weitnauer et al., 2001). Type 1 Polyketides were reported to be involved in antibiotic synthesis such as erythromycin and oxytetracycline (Yu et al., 2012), while type 3 Polyketides were known about their antibacterial and antitumor activities (Lim et al., 2016).

In this study, some environments such as Tonga trench doesn't have a signature microbial composition. However, the genus *Chewanella* was characteristic to this environment by producing a unique product (i.e. Phosphonate) which had been reported to have antibacterial activity and one famous example is fosfomycin (Metcalf & van der Donk, 2009), at the time of sampling, this species might started to fight to create its own niche. This could be studied over a course of time to detect the environmental microbial composition change over time.

Comparison of BGCs as detected by DeepBGC and antiSMASH

By using both pipelines we noticed some major differences between them and could be summarized in the following points; antiSMASH was more power in detecting the exact BGC class, on the other hand, deepBGC detects a huge amount of BGCs compared to antiSMASH. A total 776 BGCs were detected by antiSMASH from the selected projects and only 45 samples from 65 processed samples gave hits, while deepBGC detected large number of BGCs, a total 79,771 BGCs were detected from the selected projects, moreover all the 65 samples gave hits. AntiSMASH detected 26 different BGC classes and the seven major classes detected in this study which collectively represent about 80% of the total detected BGCs classes were NRPS (23%), bacteriocin (15%), NRPS-like (10%), terpene (10%), sactipeptide (9%), arylpolyene (7%) and siderophore (5%). On the other hand, deepBGC assigned BGCs classes to only around 20% of hits (15,714 hits) as follows; 39% of hits was Polyketide, 18% RiPP, 18% Saccharide, 10% others, 7% NRP, 4% Terpene, 2% Polyketide-Terpene, 1% NRP-Polyketide and 1% Saccharide-Terpene. Although deepBGC could annotate only six classes (i.e. Alkaloid, NRP, Polyketide, RiPP, Saccharide and Terpene), however it has a major advantage as it could assign product activity to each single detected BGC class with a very high coverage rate (Figure 20). In this study deepBGC assigned product activity to 96% of the hits and the results were 97% of hits have an antibacterial

activity, 1% inhibitor, 1% antibacterial-antifungal and less than 1% cytotoxic. Table 5 had a detailed comparison between antiSMASH and deepBGC in terms of the detected BGC classes with their corresponding percentage and the name and percentage of the most abundant contributed genera.

Figure 20. BGC hits detected by DeepBGC (product activity assigned) boxplot for each dataset, in relation to its assigned genus.

AMR genes profile of samples in the dataset as detected by CARD's RGI

Around 700,000 deaths yearly due to infection by resistant microbes (O'Neill (chair) J., 2014). According to ECDC, the European Center of Disease Prevention and Control, antimicrobial resistance infections cause every year around 23,000 and 25,000 deaths in the US and Europe, respectively (CDC infographic, 2019). Such figures mandate the need of novel natural products discovery with novel mechanisms of action. To reach this goal, we tried to contribute in the first part of our study through catalog the BGCs of different bacterial species from the selected metagenomes as this would help a lot in understanding the dynamics of SMs between related or different microbes and hopefully this would help. In the same context, the second major goal of this

research was to detect the antimicrobial resistance genes of the samples from the selected metagenomes, along with their mechanisms of actions and the drug classes which it confers resistance to. This also would greatly help to understand the different factors affecting the development of resistance and the possibility of spreading this resistance between closely related or even different bacterial strains through Horizontal Gene Transfer (HGT). Here we used Resistance Gene Identifier (RGI) algorithm from The Comprehensive Antibiotic Resistance Database (CARD) for AMR genes detection. From the sixty five processed samples from the different seven selected metagenomes, CARD's RGI recognized a total number of around 1216 AMR gene families which confer resistance to about 2602 drug classes by 1163 resistance mechanisms. The largest percentages of results, more than 70%, were from the samples of both gut microbiome of COVID-19 patients and water sewage (Figures 8 – 10) and they also share many of aspects as follows; among the fifty five and thirty seven different AMR gene families of gut microbiome of COVID-19 patients and water sewage, respectively, the first major two AMR gene families were resistance-nodulation-cell division (RND) antibiotic efflux pump and major facilitator superfamily (MFS) antibiotic efflux pump they represent the highest percentages; 45.75% and 45.65% of gut microbiome of COVID-19 patients and water sewage samples, respectively. Therefore, the major resistance mechanism in both samples was antibiotic efflux which represents 47.93% in gut microbiome of COVID-19 patients' samples and 45.38% in water sewage samples. Although, gut microbiome of COVID-19 patients' samples comes in the first place in the number of the detected drug classes, 1378 drugs of different 32 classes, and water sewage comes next with 1031 drugs from different 29 classes, they were also relatively similar in terms of kind of drug classes. The first three major drug classes detected in both samples were tetracycline antibiotic, fluoroquinolone antibiotic and penam which collectively represent around 35.92% and 36.95% of all detected drug classes from gut microbiome of COVID-19 patients and water sewage samples, respectively.

The rest of samples from the other projects show also similar results with a very low abundance compared to gut microbiome of COVID-19 patients' samples and water sewage samples. Antibiotic efflux pump AMR gene families both RND and MFS also represent the highest percentages; they represent 58.33% & 19.67% of all detected AMR gene families of Osaka and skin AD samples, respectively. However, there was

a slight difference in the order of resistances mechanisms between them, antibiotic efflux comes in the first place and represents 51.61% of the three detected mechanisms of Osaka samples while antibiotic inactivation comes first by 35% and antibiotic efflux was in the third place by 23.33% from the five different detected resistance mechanisms of skin AD patients samples. Moreover, the major drug class detected in both samples was aminoglycoside antibiotic by 14.85% and 12.99% from skin AD and Osaka samples, respectively. The results of the last samples from the rest three metagenomes, Nose, Tonga and Mangrove, represent the smallest fraction of all results. Their combined results were less than 1% compared to the rest of results. No perfect hits were detected and only few strict hits were reported as follows, 3 hits, 1 hit and 1 hit from Nose, Tonga and Mangrove projects, respectively.

The effect of antibiotic use in AMR genes transfer between microbial communities

We tried to find a connection between the detected BGCs in the first part of our study and AMR genes detected by CARD. Although, the ten samples of skin AD patients were the first among the rest metagenomes in terms of the number of detected BGCs, 272 BGCs from 19 different classes, it comes in the 4th place in terms of the normalized percentage of detected AMR genes after the samples of gut microbiome of COVID-19 patients, Sewage and Osaka metagenomes. Therefore, another reason other than the degree of microbial diversity should be responsible for putting the microbial communities under stress and bush them to share their resistance genes horizontally. This might be the reason behind the high abundance of the detected AMR genes, around 70%, from gut microbiome of COVID-19 patients and water sewage samples. To eliminate the effect of sample size on results, this percentage was normalized to the total number of bases per each project. Results show a degree of similarity between both samples, this might be due to a common factor drives their respective bacterial communities toward sharing their AMR genes. Microbial communities of both gut microbiome of COVID-19 patients' samples and water sewage sample exposed to abroad spectrum of antibiotics from different classes with different doses, such common factor would be of a great impact on the development of a huge number of highly resistant bacterial strains. In the next part, we will try to highlight some important results from both environments and trying to relate this to the recent researches.

AMR genes in gut microbiome of COVID-19 patients' samples

Regardless the fact that COVID-19 is a viral infection, many people around the world in low, middle and also high-income countries think that the use of antibiotics would help in the treatment and/or prevention of infection. In the same context and according to the European division of WHO, there were results of behavioural insight research from nine European countries prove that the antibiotic use increasing along with cases throughout the pandemic, around 79 – 96% of those taking antibiotics, were reported not infected with COVID-19 but they believe that the use of antibiotics is the proper preventive action. Moreover, results show that 75% of COVID-19 patients used antibiotics while only 15% of them develop bacterial co-infection and could need antibiotics (WHO, 2020). On the other hand, in Italy when the pandemic strikes, according to Dr Nino Berdzuli, Director of WHO/Europe's Division of Country Health Programmes, they gave COVID-19 patients broad spectrum antibiotics such as cephalosporins and azithromycin, this was the routine treatment of community-acquired pneumonia cases. To date, azithromycin still used as the first choice antibiotic in such cases worldwide on the basis of its immunomodulatory action. Here we reported that about 15% AMR genes confer resistance to cephalosporin and macrolide antibiotic drugs classes, azithromycin belongs to macrolides, and this would show a direct relation between the use of antibiotics and the development of antimicrobial resistance. A recent recovery trial in the UK published on 14 December, 2020 shows that azithromycin with no benefit to patients hospitalized with COVID-19. In this trial a total 2582 patients taking azithromycin were compared to 5182 patients randomized to the usual care alone (Horby et al., 2020). The situation is even worse, a study published on March, 2020 conducted in intensive care units from 88 countries on total 15165 COVID-19 patients, showed that 70% of them received antibiotics, at least one, for treatment or even prophylaxis purposes where only 54% of them had proven bacterial co-infection (Vincent JL et al., 2020). The wide use of biocidal agents as disinfectants in non-clinical, would be another possible threat. It has been reported that even the low exposure to these agents leads to the selection of drug resistance microorganisms, particularly gram negative bacteria (Kampf G., 2018). Our results showed that gut microbiome of COVID-19 patients' samples were the first among the selected metagenomes in terms of number of detected hits with total 544 hits. Moreover, it has the highest percentage of detected AMR gene families (48.26%) of different 55 families

which confer resistance to around 51.51% of drugs from different 37 classes by six different mechanisms. This might explain the reported fatal co-infection by exceptionally antibiotic resistance bacteria in gut microbiome of COVID-19 patients (Sharifipour E. et al, 2020). There should be a strict regulation from the health authorities around the world to avoid using antibiotics in cases where no sign of bacterial co-infection.

AMR genes in water sewage

Although, spread of antimicrobial resistance is a global clinical concern, this issue is not limited to the clinic. Antibiotics used by humans will at the end of the day end up in sewage, therefore waste water considered one of the biggest reservoir of antibiotics, AMR genes and bacterial from diverse sources and waste water treatment plants are usually one of the main sources of antibiotic-resistance bacteria and AMR genes spread into the environment (Rizzo, L. et al., 2013). In our study, we reported 466 hits from all samples of water sewage, 96 of hits were perfect and 370 were strict hits. This represents 23.37% of the detected AMR gene families from different 37 families which confer resistance to 22.42% drugs from different 29 classes with different 6 resistance mechanisms. Waste water considered a hotspot of spreading resistance genes not only between closely related bacterial strains but also this could be happen between phylogenetically distant strains (Jiang, X. et al, 2017) this might be due to the selection pressure caused by pollutant compounds such as heavy metals, antimicrobial agents, biocides and drugs which could promote horizontal gene transfer (Aminov, R.I, 2011). Such selection pressure is a significant issue in the presence and spreading of AMR genes in sewage. Whenever there is a selection pressure for antimicrobial resistance bacteria, they overgrow the sensitive ones and they can share their resistance genes, which are usually included in mobile genetic elements (MGE), through one of the three major mechanism, transformation, transduction and conjugation (Karkman A et al., 2018). We reported here relatively big numbers of AMR gene families, 55 from gut microbiome of COVID-19 patients' samples and 37 from water sewage, which confer resistance to many drug classes, 37 and 29, from gut microbiome of COVID-19 patients' samples and water sewage samples, respectively. This could be explained by the fact that each MGE usually contains AMR genes for more than on antimicrobial compound. Therefore, AMR gene could be selected by a wide array of antibiotics which is the case in both environments. Moreover, the same

MGE might also contain AMR gene for heavy metal or disinfectant, this can also lead to a selection pressure for transferring antibiotic resistance between different bacterial species (Karkman A et al., 2017). Waste water contributed the most in transmission of AMR genes. Results from 63 studies, published between 2009 and 2019, were reviewed elsewhere, confirming the presence of wide range of AMR genes and antibiotic-resistance bacteria in waste water around the world (Fouz N. et al., 2020).

Antibiotic efflux resistance mechanism in gut microbiome of COVID-19 patients' and water sewage samples

In bacterial drug efflux pumps have many functions other than their key role in drug resistance and there are escalating number of multiple drug efflux pumps reported from bacteria isolated from different ecological samples (Li X-Z and Nikaido H, Drugs, 2009). In the current study, antibiotic efflux was the major detected resistance mechanism from both gut microbiome of COVID-19 patients' and water sewage samples. It comes in the first place by 47.93% and 45.38% of the total detected resistance mechanisms from gut microbiome of COVID-19 patients and water sewage samples, respectively. It has been reported that efflux mediate drug resistance usually acts concurrently with other resistance mechanisms which reflects higher resistance to abroad spectrum of drugs. On the other hand, expression of drug pumps usually induced by many molecules such as antimicrobial agents, bile salts and biocides (Li X-Z and Nikaido H, Drugs, 2009), coupled with the fact that resistance genes usually present on plasmids and other mobile genetic elements, the possibilities of their induction and transfer between other bacteria in both gut microbiome of COVID-19 patients and waste water are very high. It has been reported that aminoglycosides and macrolides induce the expression of MexXY efflux pumps in *P. aeruginosa* (Jeannot K et al., 2005). Fluoroquinolones were also responsible of induction the expression of both AcrAB and PatAB pumps in *S. pneumonia* (Marrer E et al., 2006).

Chapter 5: Conclusions and Future Perspectives

Conclusions

This study has two main objectives. Firstly, the assembled contigs were investigated by two major distinct computational methods, namely antiSMASH and deepBGC methods. A comparative study was performed to determine BGCs present in each of the included samples, as well as comparing their taxonomic differences. Secondly, the assembled contigs were also analyzed to determine AMR genes present in each samples by using RGI algorithm which is a part of CARD. A total number of sixty five samples pertaining to seven selected metagenomic / metatranscriptomic projects were assembled, a total 1,139,543,039 reads were filtered to 1,100,630,009 filtered reads and the obtained reads and assembled contigs (4,325,515 contigs) were subjected to taxonomic assignment. All assemblies were then investigated by two different computational tools in addition to CARD, the first tool was antiSMASH, the second tool was deepBGC and finally we used CARD to determine AMR genes in each ecosystem.

To determine BGC content in each environment, our first goal, both computational tools, antiSMASH and deepBGC, were run in parallel for BGC mining. Although both tools were complementary to each other, however, there were major differences between them, generally in terms of total number of the detected BGCs and the number of annotated BGC classes. AntiSMASH detected only 776 BGCs which represents less than 1% of the total number of detected BGCs by deepBGC (79,771 BGCs). However, antiSMASH showed a higher accuracy in detecting the exact classes of BGCs and a higher annotation level. In this study antiSMASH annotated 26 different classes of BGCs compared to only 6 fixed classes annotated by deepBGC (i.e. Alkaloid, NRP, Polyketide, RiPP, Saccharide and Terpene) in addition to one extra unknown class named “other”. A major advantage of deepBGC was its ability to assign product activity to more than 95% of hits regardless the fact that only 20% of hits were got BGC class annotation by deepBGC. The majority of product activities assigned by deepBGC were 97% antibacterial, 1% inhibitor, 1% antibacterial-antifungal and less than 1% cytotoxic. For more detailed comparison between antiSMASH and deepBGC in terms of the

detected BGC classes with their corresponding percentage and the name and percentage of the most abundant contributed genera, see (Table 5).

The taxonomical assignment were of great impact to understand the dynamics of SM production and the differences of BGCs classes detected from the different environments. We clustered all samples based on their relative abundance of taxa and their microbial composition by both PCA and t-SNE tools (Figures 5 & 6). Some samples from the same projects had a characteristic relative microbial abundance so they appeared nicely separated from the rest of samples such as the samples from both Osaka bay project which dominated mainly by *Pseudomonas* and water sewage project which had a very high relative abundance of and *Streptococcus.* Although, these two environments dominated mainly be one genus and it was expected to have a slightly low range of BGCs and if there a unique class of BGCs will be produced by other predominated genera to protect their niche, however, *Pseudomonas* from Osaka project produced a class of BGCs called N-γ-acetylglutaminyl glutamine 1-amide (NAGGN), has a role in bacterial cell survival (Matthias Kurz et al., 2010) and it was not detected elsewhere from any of the selected samples from all other projects, *Streptococcus* also from water sewage project produced Ras-RiPP, has a role in quorum sensing, which also was not detected in any other samples. Such examples are good evidence that, also microorganisms have characteristic behavior, however, it might behave in a different ways under different environmental conditions. This needs further investigations of such environments over a course of time to see how their behavior changes over time, could be a clear limitation of this study.

On the other hand, we also expect to detect unique classes from some species which present in a very low abundance in some environments, here we reported two cases. In Tonga trench project antiSMASH detected a BGC class called phosphonate which was belonging to a genus called *Shewanella* which existed in a very low relative abundance compared to *Cutibacterium*. The second example was from the water sewage project, BGC class called phenazine was detected by antiSMASH and it was produced by *E. coli* which exhibits a low relative abundance compared to *Streptococcus.*

We also noticed that the samples which had a large extent of variability (i.e. sex, age and illness state) due to the nature of their environments, such as microbiome samples of patients in two projects (COVID-19 & Atopic Dermatitis), gave the most

variable number of BGC classes detected by antiSMASH, where 19 different classes detected in skin microbiome of AD patients and 16 different classes detected in gut microbiome of COVID-19 patients' samples while the third place in terms of total number of detected classes went to the water sewage project with 12 different classes detected and this could be also due to the same reason. AntiSMASH did not detect more than 10 different classes in the rest of projects.

The second goal of this study was to determine the AMR genes in the selected metagenomes using CARD's RGI algorithm. Due to the selection pressure on the microbial communities by the wide use of antibiotics, gut microbiome of COVID-19 patients' and water sewage samples had more than 70% of the detected AMR gene families as detected by RGI. Gut microbiome of COVID-19 patients' samples came in the first place among the seven selected metagenomes by almost 50% of the total detected AMR genes, while samples of water sewage came in the second place by almost 25%. This might be a logical result of the misuse of antibiotics all over the world as the majority of people believe this could help in the prevention or treatment of the infection, as reported in many studies. In addition to the misuse of antibiotics, the wide use of disinfectants for environmental and personal hygiene was also a potential reason of spreading of antimicrobial resistance genes between different bacterial species. Under specific harsh conditions bacterial species behave in adaptive way to survive. One major mechanism, by which the resistant bacterial species would help the sensitive ones to survive is through sharing their resistance gene horizontally by well-known mechanism called horizontal gene transfer (HGT). Mobile genetic elements, transferred during HGT, such as plasmids often contain resistance genes for more than one antibiotic, moreover, in some cases the same element might contain resistance gene for a specific metal or disinfectant. Consequently, such resistance genes might be selected by the use of wide range of antibiotics, disinfectant and heavy metals. This applied to both environments (i.e. gut microbiome of COVID-19 patients and water sewage) because antibiotics and the other pharmaceutical drugs consumed by humans will eventually end up in sewage.

Recent studies conducted on COVID-19 patients globally show the improper use of antibiotics along with many cases of fatal co-infection with highly resistant bacterial strains. In our study, gut microbiome of COVID-19 patients' samples harbor bacterial species resistant to 37 different classes of antibiotics. About 15% of the

detected AMR genes were resistant to the two major antibiotic classes used in COVID-19 infection, cephalosporins and macrolides. Despite the fact that the majority of cases, even if they are diagnosed as positive COVID-19, don't need antibiotics as long as there is no sign of bacterial infection, 75% of patients take antibiotics. On the other hand, a recovery study conducted in the UK revealed that the use of azithromycin with COVID-19 patients was with no effect compared to patients with same conditions randomized on the routine treatment without azithromycin.

Overall, diverse environments harbor different microbial composition with dissimilar relative abundance of taxa and this leads to the presence of a wide variety of secondary metabolites in each environments in addition to the presence of a wide range of AMR genes in environments under specific selection pressure such as gut microbiome of COVID-19 patients' and water sewage samples. Environments with high microbial diversity such as host microbiome (i.e. skin AD & gut COVID-19) harbor large percentage of BGCs, maybe due to the arms race between co-existing microorganisms. Both antiSMASH and deepBGC complemented each other to get a clearer picture about the nature of different environments in terms of the relative microbial abundance and their corresponding BGCs content. In addition to the degree of microbial diversity, environments under specific selection pressure by antibiotics, disinfectants and heavy metals, had the biggest percentages of AMR genes. COVID-19 and water sewage harbor more than 70% of AMR genes detected by CARD's RGI.

Future Perspectives

There is no question that there is an escalating interest to investigate biosynthetic pathways to discover new natural products. Environments with rich microbial diversity such as host microbiome and marine ecosystems should be thoroughly mined for biosynthetic gene clusters and antimicrobial resistant genes using different computational tools in order to find explanation on how novel secondary metabolites are assembled and which microorganisms carry AMR genes and to what extent they are mobile. The information in this study will be of great value to other researchers who interested in either isolation of natural products or studying the antimicrobial resistance mechanisms. In addition to their therapeutic use, understanding the dynamics of secondary metabolites is crucial for studying different microbial populations and their effects on substance turnover.

On the other hand, a time dimension could be a major limitation to this study, sampling over a course of time is critical to clearly understand the dynamics in each ecosystem. Few cases of interest were reported, we detect many SMs with antibacterial activity belonging to some genera, present in a relatively low abundance in highly diverse ecosystems, such as *Pseudomonas* from Osaka and *Streptococcus* from sewage water, they might be under stress and were trying to fight to create their own niche by producing their own weapons (i.e. SMs) at the time of sampling but we do not know how the situation could be changed over time. Moreover, the use of antibiotics and other factors such as disinfectants shift microbial populations toward sharing their resistance genes. Therefore, monitoring microbial environments over a course of time is very crucial to understand the microbial behavior under different conditions such as high competition and other stress environmental conditions such as antibiotics, disinfectants and heavy metals.

Many evidences suggested that, the misuse of antibiotics has a direct contribution to the global widespread of antibiotic resistance. In this study, samples from gut microbiome of COVID-19 patients from Hong Kong showed very interesting results, it harbors the largest percentage of AMR genes, more than 50% of the detected AMR genes in all datasets. Many factors might be contributed to such results, it would be very important to compare our results with COVID-19 patients' results from different places around the world, this would clearly unleash the role of the different

environmental factors contributed to the escalating burden of antibiotic resistance among COVID-19 patients.

In the near future, we should see a new era of development of bioinformatics tools and software based on different machine learning approaches to eliminate any current limitations and also trying to put a clear workflow optimizing the mechanism of BGC and AMR gene detection and expression of their corresponding secondary metabolites and AMR genes, respectively.

A final word to all people around the world, please keep antibiotics for patients with clear and documented signs of bacterial infection, the misuse of antibiotic will accelerate the arrival of the post-antibiotic era. Dr. Nino the director of WHO, European division said, “Everyone has a role to play as an antibiotic guardian, whether they are a parent, a prescriber or a policy-maker.”

References

Who.int. 2020. Antibiotic Resistance. [online] Available at: <https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance> [Accessed 21 August 2020].

Martens, E., & Demain, A. L. (2017). The antibiotic resistance crisis, with a focus on the United States. *The Journal of Antibiotics*, *70*(5), 520–526. https://doi.org/10.1038/ja.2017.30

Holohan, C., Van Schaeybroeck, S., Longley, D. B., & Johnston, P. G. (2013). Cancer drug resistance: An evolving paradigm. *Nature Reviews. Cancer*, *13*(10), 714–726. https://doi.org/10.1038/nrc3599

Hernando-Amado, S., Coque, T. M., Baquero, F., & Martínez, J. L. (2019). Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nature Microbiology*, *4*(9), 1432–1442. https://doi.org/10.1038/s41564-019-0503-9

Davies, J., & Ryan, K. S. (2012). Introducing the parvome: Bioactive compounds in the microbial world. *ACS Chemical Biology*, *7*(2), 252–259. https://doi.org/10.1021/cb200337h

Ruiz, B., Chávez, A., Forero, A., García-Huante, Y., Romero, A., Sánchez, M., Rocha, D., Sánchez, B., Rodríguez-Sanoja, R., Sánchez, S., & Langley, E. (2010). Production of microbial secondary metabolites: Regulation by the carbon source. *Critical Reviews in Microbiology*, *36*(2), 146–167. https://doi.org/10.3109/10408410903489576

Newman, D. J., & Cragg, G. M. (2016). Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products*, *79*(3), 629–661. https://doi.org/10.1021/acs.jnatprod.5b01055

Newman, D. J., & Cragg, G. M. (2020). Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products*, *83*(3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285

Martin, J. F. (1992). Clusters of genes for the biosynthesis of antibiotics: Regulatory genes and overproduction of pharmaceuticals. *Journal of Industrial Microbiology*, *9*(2), 73–90. https://doi.org/10.1007/BF01569737

Keller, N. P., Turner, G., & Bennett, J. W. (2005). Fungal secondary metabolism—From biochemistry to genomics. *Nature Reviews. Microbiology*, *3*(12), 937–947. https://doi.org/10.1038/nrmicro1286

Medema, M. H., & Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nature Chemical Biology*, *11*(9), 639–648. https://doi.org/10.1038/nchembio.1884

Weber, T., & Kim, H. U. (2016). The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology*, *1*(2), 69–79. https://doi.org/10.1016/j.synbio.2015.12.002

Ayuso-Sacido, A., & Genilloud, O. (2005). New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: Detection and distribution of these biosynthetic gene sequences in major taxonomic groups. *Microbial Ecology*, *49*(1), 10–24. https://doi.org/10.1007/s00248-004-0249-6

Katz, L., & Baltz, R. H. (2016). Natural product discovery: Past, present, and future. *Journal of Industrial Microbiology & Biotechnology*, *43*(2–3), 155–176. https://doi.org/10.1007/s10295-015-1723-5

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., & Weber, T. (2019). antiSMASH 5.0: Updates to the secondary metabolite genome

mining pipeline. *Nucleic Acids Research*, *47*(W1), W81–W87. https://doi.org/10.1093/nar/gkz310

Ren, H., Shi, C., & Zhao, H. (2020). Computational Tools for Discovering and Engineering Natural Product Biosynthetic Pathways. *IScience*, *23*(1), 100795. https://doi.org/10.1016/j.isci.2019.100795

Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., Suarez Duran, H. G., de Los Santos, E. L. C., Kim, H. U., Nave, M., Dickschat, J. S., Mitchell, D. A., Shelest, E., Breitling, R., Takano, E., Lee, S. Y., Weber, T., & Medema, M. H. (2017). AntiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research*, *45*(W1), W36–W41. https://doi.org/10.1093/nar/gkx319

Starcevic, A., Zucko, J., Simunkovic, J., Long, P. F., Cullum, J., & Hranueli, D. (2008). 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*(21), 6882–6892. https://doi.org/10.1093/nar/gkn685

Hannigan, G. D., Prihoda, D., Palicka, A., Soukup, J., Klempir, O., Rampula, L., Durcak, J., Wurst, M., Kotowski, J., Chang, D., Wang, R., Piizzi, G., Temesi, G., Hazuda, D. J., Woelk, C. H., & Bitton, D. A. (2019). A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, *47*(18), e110. https://doi.org/10.1093/nar/gkz654

Chen, R., Wong, H. L., & Burns, B. P. (2019). New Approaches to Detect Biosynthetic Gene Clusters in the Environment. *Medicines (Basel, Switzerland)*, *6*(1). https://doi.org/10.3390/medicines6010032

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of Bacteriology*, *194*(16), 4151–4160. https://doi.org/10.1128/JB.00345-12

Luo, Y., Cobb, R. E., & Zhao, H. (2014). Recent advances in natural product discovery. *Current Opinion in Biotechnology*, *30*, 230–237. https://doi.org/10.1016/j.copbio.2014.09.002

Li, J. W.-H., & Vederas, J. C. (2009). Drug discovery and natural products: End of an era or an endless frontier? *Science (New York, N.Y.)*, *325*(5937), 161–165. https://doi.org/10.1126/science.1168243

Jensen, P. R. (2016). Natural Products and the Gene Cluster Revolution. *Trends in Microbiology*, *24*(12), 968–977. https://doi.org/10.1016/j.tim.2016.07.006

Ghurye, J. S., Cepeda-Espinoza, V., & Pop, M. (2016). Metagenomic Assembly: Overview, Challenges and Applications. *The Yale Journal of Biology and Medicine*, *89*(3), 353–362.

Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics: TIG*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Pootakham, W., Mhuantong, W., Yoocha, T., Putchim, L., Sonthirod, C., Naktang, C., Thongtham, N., & Tangphatsornruang, S. (2017). High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Scientific Reports*, *7*(1), 2774. https://doi.org/10.1038/s41598-017-03139-4

Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., & Snyder, M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nature Biotechnology*, *34*(1), 64–69. https://doi.org/10.1038/nbt.3416

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., Waggott, D., Utiramerur, S., Hou, Y., Smith, K. S., Montgomery, S. B., Wheeler, M., Buchan, J. G., Lambert, C. C., Eng, K. S., Hickey, L., Korlach, J., Ford, J., & Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, *20*(1), 159–163. https://doi.org/10.1038/gim.2017.86

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*, *180*(18), 4765–4774. https://doi.org/10.1128/JB.180.18.4765-4774.1998

Wilson, M. C., & Piel, J. (2013). Metagenomic approaches for exploiting uncultivated bacteria as a resource for novel biosynthetic enzymology. *Chemistry & Biology*, *20*(5), 636–647. https://doi.org/10.1016/j.chembiol.2013.04.011

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., & Weber, T. (2019). antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research*, *47*(W1), W81–W87. https://doi.org/10.1093/nar/gkz310

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research*, *45*(W1), W55–W63. https://doi.org/10.1093/nar/gkx305

Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, *10*(6), 402–415. https://doi.org/10.2174/138920209789177575

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315–1316. https://doi.org/10.1038/nbt1004-1315

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279-285. https://doi.org/10.1093/nar/gkv1344

Hochreiter, Sepp, Heusel, M., & Obermayer, K. (2007). Fast model-based protein homology detection without alignment. *Bioinformatics (Oxford, England)*, *23*(14), 1728–1736. https://doi.org/10.1093/bioinformatics/btm247

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Alcock, B. P., Raphenya, A. R., Lau, T., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., … McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic acids research, 48(D1), D517–D525. https://doi.org/10.1093/nar/gkz935

Rizzo, L. et al. (2013) Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: A review. Sci. Total Environ. 447, 345-360

Vincent JL, Sakr Y, Singer M, Martin-Loeches I, Machado FR, Marshall JC, et al.; EPIC III Investigators. Prevalence and outcomes of infection among patients in intensive care units in 2017. JAMA. 2020 Mar 24;323(15):1478–87. http://dx.doi.org/10.1001/jama.2020.2717 pmid: 32207816)

Kampf G. Biocidal agents used for disinfection can enhance antibiotic resistance in gram-negative species. Antibiotics (Basel). 2018 Dec 14;7(4):110.)

Jiang, X.; Ellabaan, M.M.H.; Charusanti, P.; Munck, C.; Blin, K.; Tong, Y.; Weber, T.; Sommer, M.O.A.; Lee, S.Y.

Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. Nat. Commun. 2017, 8, 15784.

Karkman A, Do TT, Walsh F, Virta MPJ. Antibiotic-Resistance Genes in Waste Water. Trends Microbiol. 2018 Mar;26(3):220-228. doi: 10.1016/j.tim.2017.09.005. Epub 2017 Oct 13. PMID: 29033338.

Fouz N, Pangesti KNA, Yasir M, Al-Malki AL, Azhar EI, Hill-Cawthorne GA, Abd El Ghany M. The Contribution of Wastewater to the Transmission of Antimicrobial Resistance in the Environment: Implications of Mass Gathering Settings. Trop Med Infect Dis. 2020 Feb 25;5(1):33. doi: 10.3390/tropicalmed5010033. PMID: 32106595; PMCID: PMC7157536.

Li XZ, Nikaido H. Efflux-mediated drug resistance in bacteria: an update. Drugs. 2009 Aug 20;69(12):1555-623. doi: 10.2165/11317030-000000000-00000. PMID: 19678712; PMCID: PMC2847397.

Jeannot K, Sobel ML, El Garch F, Poole K, Plésiat P. Induction of the MexXY efflux pump in Pseudomonas aeruginosa is dependent on drug-ribosome interaction. J Bacteriol. 2005 Aug;187(15):5341-6. doi: 10.1128/JB.187.15.5341-5346.2005. PMID: 16030228; PMCID: PMC1196038.

Marrer E, Satoh AT, Johnson MM, Piddock LJ, Page MG. Global transcriptome analysis of the responses of a fluoroquinolone-resistant Streptococcus pneumoniae mutant and its parent to ciprofloxacin. Antimicrob Agents Chemother. 2006

Jan;50(1):269-78. doi: 10.1128/AAC.50.1.269-278.2006. PMID: 16377697; PMCID: PMC1346767.

Bengtsson-Palme, J., Kristiansson, E., & Larsson, D. (2018). Environmental factors influencing the development and spread of antibiotic resistance. FEMS microbiology reviews, 42(1), fux053. https://doi.org/10.1093/femsre/fux053

Zuo, T., Zhang, F., Lui, G., Yeoh, Y. K., Li, A., Zhan, H., Wan, Y., Chung, A., Cheung, C. P., Chen, N., Lai, C., Chen, Z., Tso, E., Fung, K., Chan, V., Ling, L., Joynt, G., Hui, D., Chan, F., Chan, P., … Ng, S. C. (2020). Alterations in Gut Microbiota of Patients With COVID-19 During Time of Hospitalization. Gastroenterology, 159(3), 944–955.e8. https://doi.org/10.1053/j.gastro.2020.05.048